Treball de Fi de Grau

Grau en Enginyeria en Tecnologies Industrials

Auditoria, preparació i anàlisi de dades usant tècniques de Data Mining en una empresa d'eficiència energètica

MEMÒRIA

Autor: Bernat Serra Frigola
Director: Lluís Talavera Mendez

Convocatòria: Juny 2016

Escola Tècnica Superior
d'Enginyeria Industrial de Barcelona

Resum

L'objectiu d'aquest projectes es avaluar la viabilitat i l'impacte de la implementació d'un entorn de processos analítics, en l'empresa de serveis d'eficiència energètica Enertika.

La realització d'aquesta avaluació es fa a partir d'una prova de concepte, és a dir, implementar un entorn de processos analítics bàsic, en un dels projectes de l'empresa, per extrapolar el potencial impacte i viabilitat de la implementació d'un entorn més elaborat per tots els projectes i departaments de l'empresa.

Inicialment es fa un estudi de l'estat de l'art de les metodologies d'aplicació de projectes de Data Mining (procés de descobrir patrons i/o relacions en grans volums de dades) i Data Warehousing (procés de creació d'una estructura per emmagatzemar dades per facilitar-ne l'anàlisi). A partir dels coneixements adquirits i els requisits, temporals i econòmics, imposats desenvolupar i adaptar una metodologia

Per aplicar la prova de concepte s'ha desenvolupat una metodologia a seguir. Alguns dels seus trets diferencials, respecte les metodologies més usades, són els següents: a priori no estableix uns objectius molt específics, el procés és més exploratori, és flexible respecte l'abast de l'anàlisi, requereix una inversió molt menor. La metodologia proposada ha sigut capaç d'acomplir amb els objectius i requeriments establerts.

Com a resultat d'aplicar la prova de concepte s'ha obtingut resultats que avalen la implementació d'un entorn de processos analítics en els altres projectes de l'empresa, a la vegada ha demostrat ser flexible amb l'abast de la seva execució, que dependrà dels recursos que es destinin al realitzar els anàlisis.

El resultat de l'objectiu principal d'aquest projecte és positiu, per tant queda demostrada la viabilitat de crear i mantenir un entorn de processos analítics, utilitzant eines *free and open source* i amb uns recursos econòmics i humans limitats

Sumari

1. Glossari

DM Data Mining

KDD Knowledge Discovery in Database

DW Data Warehouse

EB Estació Base

WTB Wattabit

FC Free-Cooling

CFE Comissió Federal d'Electricitat

AA Aire Condicionat

BBDD Base de Dades

ETK Enertika

TEF Telefònica

UMTS Sistema Universal de Telecomunicacions Mòbils (*Universal Mobile Telecommunication System)*

PLC Controlador lògic programable (*Programmable Logic Controller*)

2. Introducció

2.1. Origen del projecte i motivació

Enertika és una empresa de serveis d'eficiència energètica. El seu negoci es centra en el desenvolupament de projectes per reduir el consum energètic, d'empreses o administracions públiques, on el retorn de les inversions realitzades s'obté a través dels estalvis generats. La majoria dels projectes executats es gestionen remotament gràcies al monitoratge, via Internet, de les variables de funcionament energètic més representatives de cada projecte. Com a conseqüència del monitoratge, l'empresa disposa d'una gran volum de dades que es generen i emmagatzemen. Aquest fet ha motivat que, en la empresa, hi hagi un interès creixent en la possible aplicació de tècniques d'anàlisi de manera regular. Aquesta situació s'ha vist com un bon punt de partida per la realització d'un projecte per estudiar la viabilitat de crear un entorn de treball per portar a terme processos analítics. Com a resultat d'aquest projecte s'espera obtenir una valoració dels requeriments, canvis i implementacions necessaris, a partir de la qual avaluar l'impacte que podria suposar una implementació similar a nivell global dins de l'empresa.

2.2. Requeriments previs

Els requeriments estipulats per a la realització d'aquest projecte són bàsicament dos.

- **Baix cost econòmic**. Aquest requeriment ve imposat per la manca de recursos econòmics destinats a la realització del projecte. Té una forta influència, sobretot, en l'elecció del *software* necessari per al tractament i integració de dades, que haurà de ser llicenciat de manera gratuïta.

- **Horitzó temporal definit**. Aquest requeriment ve imposat per la voluntat d'obtenir resultats en un període d'entre 3 i 5 mesos. Té una important influència en l'abast i objectius del projecte, ja que s'hi hauran d'adaptar. Per tant s'haurà de posposar l'aspiració de realitzar models d'anàlisi complexes.

2.3. Objectius del projecte

- Estudiar la viabilitat i l'impacte de la implementació d'un entorn de processos analítics, per poder realitzar diversos tipus d'anàlisi de manera regular, a les dades que es generen en l'operativa de l'empresa de serveis d'eficiència energètica Enertika.

    - Estudiar l'estat de l'art de les metodologies de Data Mining i anàlisi de dades.
    - Desenvolupar, seleccionar i/o adaptar una metodologia per a l'execució del projecte, és a dir, per a l'estudi, preparació i anàlisis de dades, que a la vegada permeti complir els requeriments temporals establerts.
    - Estudiar i seleccionar les eines de tractament de dades a emprar per realitzar el projecte complint amb els requeriments previs.
    - Aplicar la metodologia proposada amb les eines elegides per tal de fer-ne una prova de concepte, a partir de la qual poder avaluar la viabilitat d'implementació.

Objectius personals

- Adquirir i comprendre els conceptes relatius al camp del Data Mining i l'anàlisi de dades.
- Comprendre tots els passos previs necessaris per la realització d'un procés de Data Mining.

2.4. Abast del projecte

En aquest projecte es vol arribar a demostrar, que tot i l'existència d'uns requeriments temporals i econòmics, és possible, com a mínim, dur a terme un anàlisi bàsic amb el que es pugui obtenir algun tipus de resultat útil que justifiqui la implementació d'un entorn analític en l'empresa.

2.5. Estructura de la memòria

El projecte s'ha estructurat en un seguit de capítols que es comenten a continuació.

- **Estat de l'art**. En aquest capítol es fa un resum dels conceptes Data Mining i Data Warehousing, per introduir al lector en aquest camp. Seguidament es presenten i discuteixen les metodologies que més s'usen o que són d'interès especial per la realització del projecte.

- **Metodologia a aplicar**. En aquest capítol s'exposa la metodologia proposada, amb l'explicació de totes les fases, accions i documents ideats per assolir els objectius del projecte amb èxit. Els següents capítols contenen la prova de concepte de la metodologia exposada i es titulen a partir de les fases proposades

- **Comprensió del negoci**. En aquest capítol s'explica la tipologia de l'empresa, el projecte de l'empresa sobre el que s'actuarà i el seu funcionament. La intenció d'aquest capítol és contextualitzar el camp d'aplicació de la metodologia.

- **Auditoria i comprensió de dades**. En aquest capítol s'exposa la tipologia, naturalesa i localització de les dades que genera l'empresa. També conté un document on es plasmen coneixements, problemàtiques i solucions detectades durant el transcurs de les dues primeres fases de la metodologia.

- **Neteja, integració i preparació preliminar**. En la part inicial d' aquest capítol s'exposen totes les accions realitzades i problemàtiques trobades durant la neteja de les dades, seguidament s'explica el procés seguit per integrar, enriquir i millorar totes les dades en una sola taula, i finalment es fa un breu resum de la taula base d'anàlisi generada.

- **Anàlisi preliminar**. En aquest capítol es mostren els anàlisis realitzats a la taula base per tal d'extreure algun tipus de coneixement contingut en les dades. També conté les conclusions i explicacions dels resultats de l'anàlisi. Els resultats extrets serveixen de precedent per demostrar el potencial de la realització d'anàlisis més extensos i/o elaborats.

- **Avaluació, recomanacions i desplegament**. En aquest capítol s'avalua la implementació de la metodologia proposada i el potencial d'estendre-la al conjunt de l'empresa, també s'exposen un seguit de recomanacions i accions de desplegament realitzades o planificades.

- **Eines usades**. En aquest capítol s'expliquen les eines avaluades que s'han triat o descartat, per la realització del projecte, tenint en compte les restriccions imposades.

3. Estat de l'art

El Data Mining és el procés de descobrir patrons i/o relacions interessants en grans volums de dades. Usualment el terme no només es refereix a l'anàlisi de les dades sinó a tot un procés integral que parteix de les dades "en brut" (*raw data*) fins a l'obtenció de coneixements útils [1]. Aquest procés integral que rep el nom de *Knowledge Discovery in Database* (KDD) [2] [3], representa el procés complet i els passos necessaris per extreure aquest coneixement. Aquests passos es poden veure en la Figura 1.

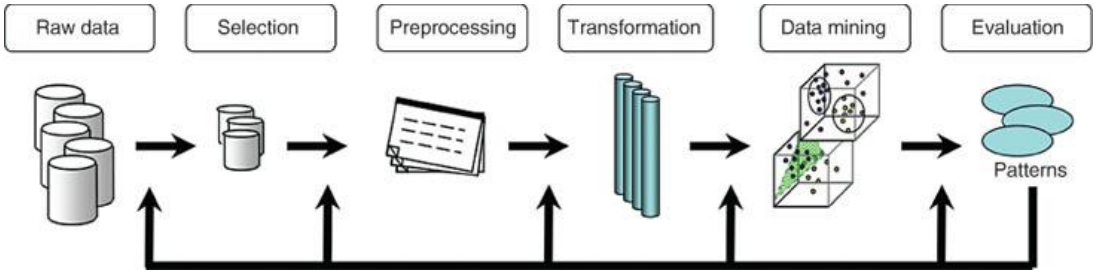


Figura 1. Esquema dels passos d'un KDD

El KDD no és una metodologia, sinó un model conceptual originat al món acadèmic, és per això que des de l'àmbit del negoci s'han proposat diverses metodologies que proposen amb detall diverses fases, establint un seguit de bones pràctiques per maximitzar les probabilitats d'èxit del projecte.

La metodologia més extensa i utilitzada és [4] CRISP-DM (*Cross Industry Standard Process for Data Mining*), que segon el portal web *kdnuggets* (portal de referència en l'anàlisi de dades) l'usen un 43% del enquestats [5]. Aquesta metodologia va ser concebuda per cinc grans companyies del sector de l'anàlisi de dades l'any 1996 i des d'aquell moment ha esdevingut un dels referents en el sector. Tot i existir alguns estàndards en una gran quantitat de projectes es fan servir metodologies pròpies, ja siguin de l'organització o de l'equip que dirigeix el DM.

D'altra banda, el Data Warehousing és el procés de dissenyar i materialitzar una estructura per emmagatzemar les dades amb el propòsit de facilitar-ne l'anàlisi. El terme Data Warehouse també s'usa per referir-se al repositori central de dades. Els dos termes DM i DW estan relacionats entre sí, ja que si les dades estan correctament emmagatzemades en un DW el procés de DM és molt més eficient i fàcil [6].

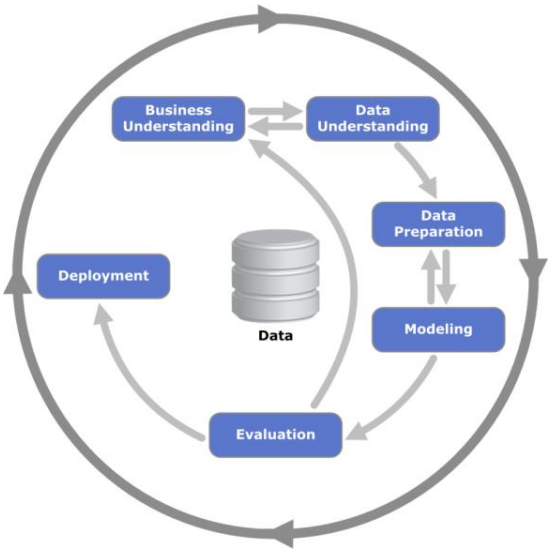


Figura 2. Diagrama de la metodologia CRISP-DM

3.1. CRISP-DM

La metodologia CRISP-DM es centra en la consecució d'objectius de negoci, en ella es separa el procés de DM en 6 parts fonamentals, seguint el diagrama de fluxos de la Figura 2.

Business Understanding

És la fase inicial del projecte, es centra en entendre els objectius i requeriments des d'una perspectiva de negoci, per posteriorment definir un pla preliminar per assolir els objectius des d'una perspectiva de data mining. D'aquesta manera s'obté una visió amplia dels recursos, riscs, restriccions...

Data Understanding

Aquesta fase s'inicia amb la recol·lecta, descripció i exploració d'informació. L'objectiu d'aquesta fase és habituar-se a la tipologia, format, etc. de les dades. S'obté una visió de la qualitat de les dades disponibles. En alguns casos es poden començar a plantejar certes hipòtesis.

Data Preparation

Al final d'aquesta etapa s'obtindrà el conjunt complet de dades (*dataset*) preparades per aplicar-hi diversos models. S'agrupen totes les dades obtingudes de diverses fonts (*Integration*) i s'hi aplica transformacions, tant per a netejar les dades com per a preparar-les pel posterior anàlisi. Aquesta fase acostuma a ser la fase que consumeix més temps del

projecte, ja que pot ser requerida més d'una vegada en funció dels model que es volen aplicar (tal i com es pot veure a la Figura 2).

Modeling

Es seleccionen diverses tècniques de modelat de dades, i es calibren els seus paràmetres per obtenir uns resultats òptims. En funció de la tècnica seleccionada es requerirà o no una nova transformació de les dades.

Evaluation

S'avalua rigorosament els processos seguits per a la construcció del model per assegurar-ne la seva qualitat analítica. També es verifica que es poden assolir els objectius definits en la primera etapa. Un dels objectius de l'etapa és determinar algun problema que no s'hagués considerat amb suficient atenció.

Deployment

Un cop s'ha obtingut un cert coneixement, gràcies a les dades, s'ha de plasmar, organitzar i presentar de la manera més útil possible per a aquell que en farà ús.

3.2. Data Warehousing

El *Data Warehouse* o magatzem de dades és una col·lecció de dades estructurades per a la consulta i anàlisi. Aquests repositoris són variants en el temps, s'actualitzen i es registren els canvis; no volàtils, la informació no es modifica ni es borra; integrats, contenen les dades de tots els sistemes operacionals de l'empresa [6].

L'objectiu d'un *Data Warehouse* és tenir tota la informació, relacionada amb una organització, en un sol lloc i preparada per aplicar-hi anàlisis que puguin ser d'utilitat a l'hora de prendre decisions.

No hi ha un únic mètode per crear un DW, però aquests passos ideats per David Walls i Mark D. Scott [7] poden servir per tenir una visió general del procés.

Determinar els objectius del negoci

En aquesta fase s'ha parlar amb els alts executius i els caps de cadascun dels departaments per tal d'entendre quins objectius tenen i sobretot quins són els factors usats per valorar el grau d'èxit que s'assoleix. Aquests factors són els anomenats KPI (*key performance indicator*), i cal tenir-los presents.

Recol·lectar i analitzar informació

Es comença per les fonts de dades que s'usen en l'actualitat per prendre decisions, s'han d'obtenir còpies de tots aquests documents i conèixer d'on provenen i com es generen. Sovint, durant el funcionament operatiu es creen documents amb informació que un cop s'han usat no se'ls dóna importància. Un dels reptes més importants d'aquesta fase és localitzar aquestes fonts d'informació i entendre perquè s'han creat, amb l'objectiu de valorar-ne la seva importància.

Identificar els processos més importants del negoci

Arribat aquest punt s'hauria de tenir una idea clara dels processos més importats del negoci, els KPI que en valoren el funcionament i com obtenir els valors d'aquests KPI.

Construir un model conceptual de dades

En aquesta fase es detalla la Taula de fets, que és la taula central de l'esquema dimensional i que conté els indicadors de funcionament del negoci. En aquesta taula cada mesura es pren intersecant cadascuna de les dimensions que la defineixen. Usualment en el DW aquestes dimensions corresponen a taules que rodegen la taula de fets, en forma d'estrella. És necessari planificar bé l'estructura que s'usarà, ja que canviar-la suposa una gran pèrdua de recursos temporals i econòmics.

Localitzar les fonts de dades i planificar la transformació de dades

Un cop es coneix la informació que es necessita s'ha d'identificar on està aquesta informació i pensar com s'ha d'integrar a l'estructura del DW. En aquesta fase cal crear un ETL (*Extract Transform and Load*) i la seva freqüència d'execució. Aquesta eina accedirà on s'emmagatzema la informació que es necessita, la transformarà de manera que sigui concordant i útil, i la carregarà al DW. Aquesta transformació és un procés molt important, pot anar des de simples transformacions de format o unitats fins a sistemes automàtics de gestió de valors no existents.

Ajustar el seguiment temporal

Un DW consumeix molt espai, ja que les dades són emmagatzemades, teòricament, per sempre. Per no generar una quantitat d'espai inassumible es pot emmagatzemar en diferents granularitats en funció de la seva antiguitat. Si aquesta granularitat es planifica és més fàcil evitar problemes posteriors amb els models analítics.

Implementar el pla

Un cop s'ha desenvolupat un pla d'actuació es podrà tenir una idea prou detallada de la quantitat de recursos i temps que seran necessaris per executar-lo.

3.3. Discussió

És important tenir present que les metodologies semblants al CRISP-DM, tenen com a objectiu principal obtenir unes fites que aportin benefici econòmic de la manera més segura possible. Això s'accentua, ja que normalment s'usa en projectes que per ser tirats endavant necessiten una gran inversió econòmica.

Per la realització d'aquest projecte no seria realista realitzar un projecte complex de Data Mining ni la creació d'un Data Warehouse tal i com s'entén, ja que la creació d'una base de dades adequada seria massa complexa i requeriria una inversió temporal massa elevada. Tot i això hi ha certes fases d'aquestes metodologies que són interessants de cara a la realització d'aquest projecte, i a la vegada poden ser molt semblants entre elles tot i provenir de metodologies amb diferents objectius. Aquestes fases són les següents:

- *Business Understanding* i *Determinar els objectius del negoci*.
- *Data Understanding* i *Recol·lectar i analitzar informació*.
- *Data Preparation* i *Localitzar les fonts de dades i planificar la transformació de dades.*

3.4. Una visió alternativa de la preparació de dades

El Data Mining també es pot enfocar des d'altres perspectives. Per exemple, el *data scientist* Richard Boire [8] exposa que solucionar un problema específic del negoci no ha de ser necessàriament l'objectiu d'un projecte de Data Mining, sinó que l'objectiu pot ser l'exploració i descobriment de dades i l'aprenentatge que se n'extreu.

Tot i la naturalesa menys definida d'aquest tipus de projecte cal seguir fixant unes directrius i passos a seguir per assegurar el seu èxit. En aquest cas es proposen quatre passos.

Preparació

Tot i que Richard Boire l'anomeni preparació no s'ha de confondre amb la preparació de dades, ja que no hi té re a veure.

En aquesta fase l'analista o el seu equip procurarà obtenir el màxim coneixement del negoci, sobretot dels aspectes que són intrínsecs del sector o negoci. Les tasques inicials inclouen

entrevistes amb responsables de les diverses àrees de l'empresa, d'aquestes es pot extreure els problemes principals amb els que es toparà, les tasques que es duen a terme a cada secció, els reptes més importants del negoci...

Auditoria de dades

En aquesta fase s'extreuen i es treballen totes les dades que siguin considerades importants, l'objectiu de l'analista és aconseguir una coneixement de les dades molt elevat de manera que les dades li siguin familiars i properes.

Un cop extretes totes les dades s'elaboren informes estandarditzats que recullen la naturalesa i qualitat de les dades. També es realitza un informe general de les dades on s'indiquen els problemes de qualitat destacats i com millorar-los.

Anàlisi preliminar

Tot i el seu nom "preliminar" és l'objectiu que es vol assolir, adquirir un coneixement bàsic a través de l'anàlisi de dades. Tot i que a mesura que el temps i els models analítics avancin s'obtindran coneixements més sofisticats. El tipus de anàlisi variarà depenent del camp en el que s'apliqui, tot i que moltes vegades implica segmentar clients, o trobar patrons de funcionament.

Recomanacions

Un cop finalitzats l'exploració i l'anàlisi s'hauran fet certs descobriments. Aquests descobriments s'han de consolidar en un document que contingui tots els coneixements que s'han extret així com confeccionar un full de ruta per dur a terme nous anàlisis.

4. Metodologia a aplicar

Es planteja seguir una metodologia molt propera a la exposada per Richard Boire (apartat 3.4), que a la vegada conté fases semblants a les metodologies CRISP-DM i DW, però que sobretot es diferencia d'elles per no tenir un objectiu inicial de negoci i ser més flexible. La justificació d'aquesta elecció ve donada per les restriccions temporals i econòmiques imposades per la realització d'aquest projecte. La metodologia proposada té com a objectiu endinsar-se en les dades de l'empresa per valorar i estudiar diversos factors, aquests són els següents:

- La qualitat de les dades generades i emmagatzemades.
- La gestió de la informació i les seves possibles millores.
- La viabilitat de seguir realitzant anàlisis de DM o la creació d'un DW.
- Marcar quin seria el camí a seguir per seguir realitzant aquest tipus d'anàlisis.

En la Figura 3 es pot observar l'esquema de la metodologia proposada. En aquest esquema s'hi mostren tres tipus d'entitats que es descriuen a continuació:

- Fases. Són els estats en què s'ha dividit el procés per millorar-ne la comprensió i incrementar les probabilitats d'èxit. Estan situades a la part esquerra de l'esquema i es simbolitzen com a rectangles.
- Accions. Són les actuacions que cal dur a terme en cada fase, per aquesta raó sempre surten d'una fase. Estan situades en la part central de l'esquema.
- Lliuraments. Són els documents que s'entreguen a la direcció de l'empresa després d'haver realitzat alguna de les fases proposades. Estan situats en la part dreta de l'esquema i separats de les altres entitats, ja que aquests són els únics productes de l'aplicació de la metodologia que arribaran a la direcció de l'organització.

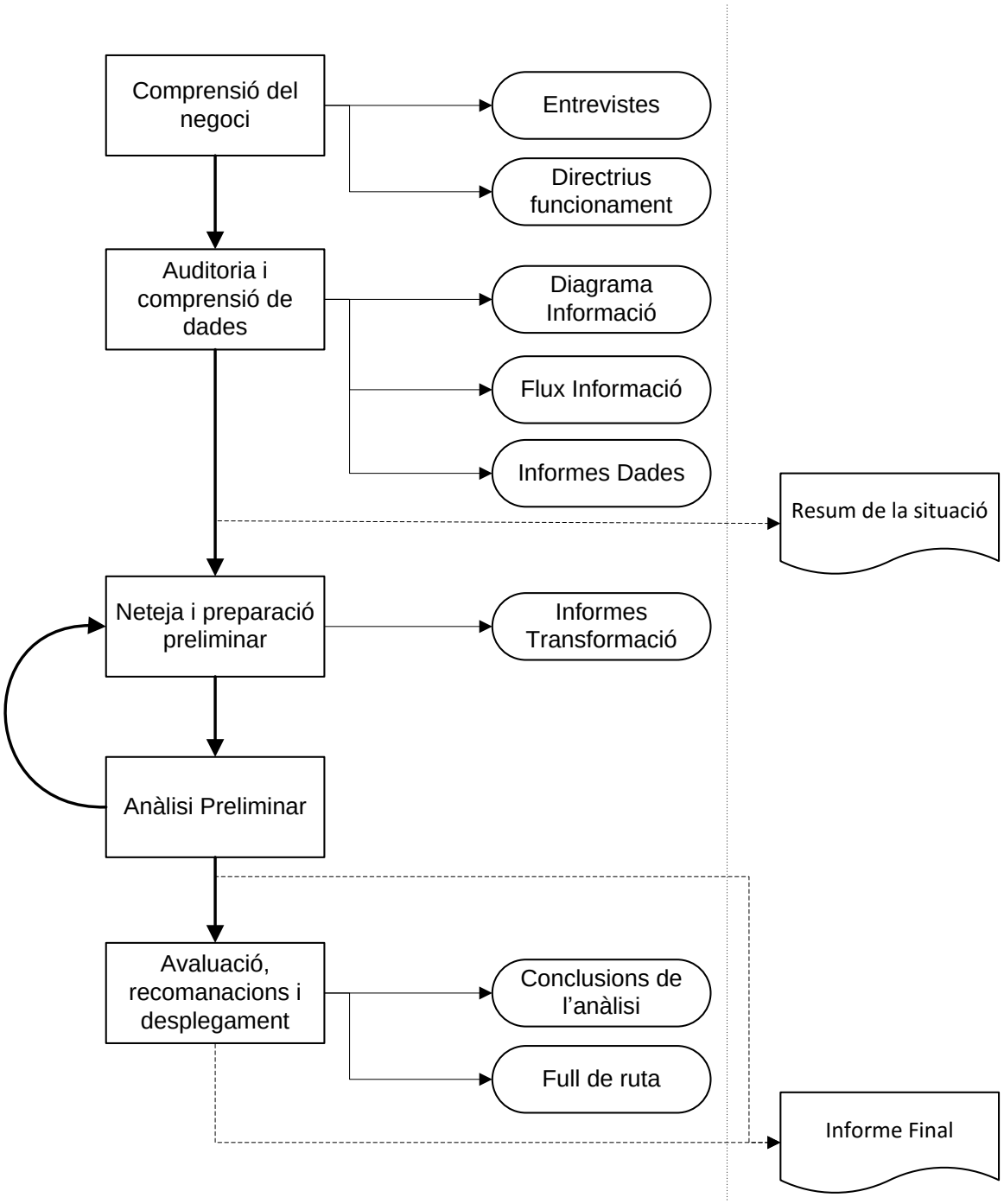


Figura 3. Diagrama de la metodologia

4.1. Comprensió del negoci

L'objectiu és obtenir un grau de coneixement del negoci suficient per entendre com funciona, quins objectius es tenen i quines dades genera el seu funcionament. Per obtenir aquest coneixement és convenient començar per tenir una visió global del negoci per després aprofundir en l'organització dels diversos departaments i també en els processos. Per aconseguir aquest coneixement és necessari concertar entrevistes. Una bona manera és seguir un ordre jeràrquic de major a menor responsabilitat, començant pels directius després els caps de departament, i seguint així fins adquirir un coneixement suficient. També és molt important llegir els documents que continguin les directrius de funcionament de l'empresa, del departament o dels processos que es realitzen. És necessari que aquestes dues fonts de coneixement es comparin de manera crítica, per conèixer, per exemple, si els protocols establerts es compleixen, si aquests han evolucionat durant el temps, etc.

Un cop acabada aquesta fase, l'analista hauria de tenir un coneixement prou detallat del funcionament del negoci, dels objectius marcats per l'empresa i dels indicadors KPI que s'usen per valorar el seu funcionament.

4.2. Auditoria i comprensió de dades

L'objectiu és tenir coneixement de tota la informació disponible, comprendre-la i valorar la seva qualitat. La qualitat de les dades [9] és una avaluació de l'aptitud de les dades per servir per un propòsit determinat en un context donat. Per avaluar la qualitat, s'han de valorar diferents aspectes, els més importants són els següents:

- Completesa, es relaciona amb el nombre d' entrades sense cap valor o amb valors incomplets.

- Exactitud, es relaciona amb mesures esbiaixades, inapropiades o que contenen errors.

- Coherència, es relaciona amb la coherència de les dades entre diferents fonts

En aquesta fase cal explorar tota la informació que es genera i s'emmagatzema dia a dia, aquesta exploració serveix per saber quina informació pot ser important, on està disponible i si pot ser útil, o no, en funció de la seva qualitat. Aquest procés es pot comparar amb un cerca en amplada, ja que la intenció és tenir en tot moment una visió global de la informació, per mica en mica endinsar-se en les diferents dades. D'aquesta manera és més fàcil mantenir una perspectiva global, i per tant disposar d'una major sensació de control i coneixement de la informació disponible.

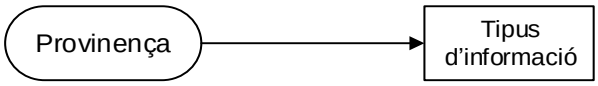


Figura 4. Notació del Diagrama d'informació existent

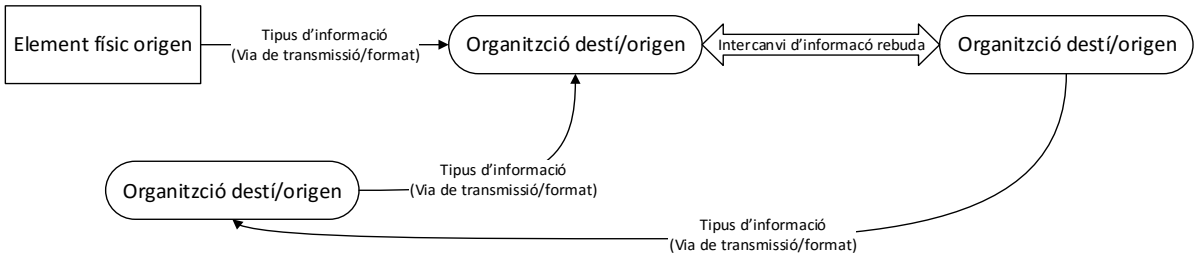


Figura 5. Notació del Diagrama de flux d'informació

4.2.1. Exploració de dades

Com a resultat de l'exploració es proposa fer una representació gràfica de la informació disponible, mitjançant un diagrama o esquema. D'aquesta manera l'analista es pot fer una idea clara de la informació de la qual disposa, ja que moltes vegades aquesta pot provenir de diversos llocs, estar entre documents de diverses índoles o estar desats en un laberint de carpetes. Aquests fets dificulten l'obtenció d'una visió sintètica de la informació que es té. En aquest diagrama s'hi farà constar el tipus d'informació i la seva provinença que pot variar en funció de l'entorn en que s'apliqui la metodologia: una organització, un servidor, una base de dades, etc. Per a la realització d'aquest esquema es proposa seguir la notació que es mostra en la Figura 4.

Un segon esquema que es proposa és la representació del flux d'informació existent. En aquest diagrama es mostra el tipus d'informació, l'origen, el destí, la via de transmissió i el format de la informació. De la mateixa manera que en l'esquema d'informació, l'origen i destí poden ser molt diversos, per exemple, un element físic, una organització, un departament, etc. Per a la realització d'aquest diagrama es proposa seguir la notació que es mostra en la Figura 5.

Després d'obtenir una idea clara de la informació disponible, cal familiaritzar-se amb les dades i valorar-ne la seva qualitat, això significa que s'ha de donar resposta a diverses preguntes sobre les dades, com per exemple: Com s'han recollit?, Perquè s'han recollit?, D'on provenen?, On s'emmagatzemen?, Quin ús se'n dona?, Qui hi té accés?, Com s'actualitzen?, etcètera. La resposta a aquestes qüestions queda plasmada en l'Inventari de dades.

METADADES	ID	Identificador del document (Ex. 00001)	Nom	Nom descriptiu de les dades compreses en el fitxer	Data	Data d'exploració
Descripció	Informació general que contenen les dades i el seu abast. (Ex. Abast: Geogràfic, Temporal…)					
Creador	Persona o entitat creadora de les dades					
Font	Origen de les dades (Ex. Sensors, Especificacions, Persones…)					
Responsable/Gestor	Persona responsable de gestionar les dades					
Funció	Ús que es dóna a les dades, a qui o a què van dirigides.					
Format	Format en que s'emmagatzemen les dades.					
Ubicació	Lloc on es troben físicament les dades. (Ex. Servidor, Cinta magnètica, Capeta , Base de dades, Taula de la base de dades…)					
Restriccions	Restriccions que limiten l'accés a les dades. Qui pot accedir a les dades					
Actualitzacions	Periodicitat d'actualització de les dades, en cas de ser actualitzades.					
Rèpliques i Còpies	Informació sobre les rèpliques i còpies de seguretat, protocol a seguir per fer-les. (Ex. On es fan, Quan es fan…)					
Fiabilitat	Possibles inconsistències conegudes en les dades. (Ex. Errors de medició, No captura durant un període, Introducció manual no fiable…)					
Comentaris	Qualsevol comentari que sigui d'utilitat per tal d'entendre i/o obtenir més informació de les dades.					

Taula 1.Metadades (Inventari de dades)

4.2.2. Inventari de dades

Per donar resposta a les qüestions sorgides en l'exploració de dades i per familiaritzar-se amb el contingut i la qualitat dels documents, s'ha de realitzar un informe. Aquest informe, anomenat inventari de dades, es realitza per a cada font d'informació[1], i en ell hi ha de quedar perfectament descrita la naturalesa, i la qualitat de la informació que conté. Després d'haver treballat amb les dades i realitzar-ne els inventaris corresponents s'haurà adquirit un grau de coneixement elevat sobre les dades a tractar.

[1] *D'ara endavant s'usarà el terme font d'informació per fer referència a cada bloc de dades corresponents a una temàtica concreta i contingudes en un mateix document.*

ESTUDI DADES				ID	Identificador del document (Ex. 00001)		Nom	Nom descriptiu de les dades compreses en el fitxer			Data	Data d'exploració
Variables	**Col.**	**Mesura**	**Escala**	**Descripció**	**nº**	**Únic**	**Tipus**	**Max**	**Min**	**Mean**	**StDev**	**Observacions**
Nom de la variable	Columna on està la variable	Unitats mesura	Escala (Ex. Nominal, categòrica…)	Descripció i observacions que ajudin a comprendre la variable.	nº de valors	nº de valors únics	Tipus de variable (Ex. Entera, Decimal, Alfanumèrica…)	**Cas numèric:** Max: Valor màxim Min: Valor mínim Mean: Mitjana St Dev: Desviació Estàndard				Observacions i informació que es consideri útil de la variable (Ex. Informació sobre la seva fiabilitat…)
								Cas alfanumèric: Max: Valor màxim de freqüència Min: Valor mínim de freqüència Mean: Mitjana de la freqüència St Dev: Desviació Estàndard de la freqüència. *(Freqüència = freqüència d'aparició de valors)*				

Taula 2. Estudi de dades (Inventari de dades)

4.2.2.1. Metadades

L'objectiu d'aquesta part és plasmar, de manera explícita, tota la informació referent al document i la seva accessibilitat, que es troba (o no) de manera implícita en el document. Per acomplir amb aquest objectiu s'ha ideat un document que es pot veure comentat en la Taula 1.

4.2.2.2. Estudi de Dades:

L'objectiu d'aquesta part és obtenir un coneixement més específic de les dades que s'emmagatzemen en el document i la qualitat de les mateixes. Per acomplir amb aquest objectiu s'ha ideat un document es pot veure comentat en el la Taula 2. S'ha de omplir la informació de la Taula 2 per a cadascuna de les variables que pugui contenir el document.

En el cas d'una variable alfanumèrica evidentment els valors màxims i mínims no estan definits, per aquesta raó es fa ús de la freqüència. La freqüència és el nombre de vegades que un valor apareix en la variable.

4.2.3. Resum de situació

Un cop acabada l'auditoria i comprensió de dades és desitjable realitzar un resum d'aquesta fase, aquest resum és el primer informe que es lliurarà a l'organització com a resultat de l'aplicació de la metodologia. L'informe ha de contenir una valoració de la situació, informació d'utilitat per a pròximes exploracions, recomanacions i qualsevol coneixement extret que l'analista cregui que ha de destacar.

Informació extreta

En aquest apartat l'analista ha de plasmar tota aquella informació continguda en les dades o metadades que li hagi semblat inusual o digne de menció.

Problemàtica

En aquest apartat l'analista/auditor ha de deixar constància de tots els processos o fets que puguin suposar un problema per a l'organització i que dificultin l'anàlisi de dades.

Solucions i recomanacions

En aquest apartat l'analista proposa solucions o recomanacions per a cadascuna de les problemàtiques trobades. Cal donar importància a la forma i contingut d'aquest apartat, ja que serà el que mirarà l'equip directiu per discutir la necessitat de l'execució de les solucions i recomanacions proposades. Per tant, és molt important que les propostes quedin escrites de manera clara, concisa i amb un llenguatge planer que pugui ser entès per algú sense coneixements de la matèria en qüestió.

Conclusions i pròxims passos

En aquest apartat s'han de plasmar les conclusions extretes durant l'auditoria i comprensió de les dades, i també s'hi comenta els pròxims passos a realitzar per seguir endavant, és a dir, amb quina informació s'aprofundirà i es seguirà estudiant i quina no.

La tria d'informació útil es fa com a últim pas del procés d'auditoria i coneixement, ja que arribats en aquest punt, l'analista té una idea prou clara de les dades que disposa. L'objectiu d'aquesta tria és ajustar l'estudi als recursos temporals i materials disponibles, tot i això cal deixar documentat la informació que s'ha seleccionat o descartat i la seva raó.

Es pot descartar informació per diverses raons, les més usuals són:

- Informació no útil: cal explicitar perquè la informació no és útil, això pot ser degut a la falta de fiabilitat o qualitat.

- Dificultat per extreure la informació: aquesta dificultat pot suposar que no sigui viable obtenir aquesta informació, per exemple, extreure informació continguda en desenes de milers de fotografies. És molt important deixar constància de perquè no és viable obtenir aquesta informació, ja que en funció d'això es podria convertir en viable per un anàlisi posterior que disposes de més recursos.

NETEJA DADES	ID	Identificador del document (Ex. 00001)	Nom	Nom descriptiu de les dades compreses en el fitxer	Data	Data d'exploració
Variables	**Neteja Aplicada**					
Nom de la variable	Transformació aplicada a la variable per tal d'eliminar o corregir-hi valors no vàlids (Ex. Extreure text de variables numèriques, Eliminar valors erronis…) O eliminació de la variable en cas que no aporti cap informació nova.					

Taula 3. Informe de neteja de dades

4.3. Neteja, integració i preparació preliminar

L'objectiu d'aquesta fase és deixar les dades suficientment netes i preparades perquè a l'hora d'aplicar-hi models analítics sigui necessària la menor quantitat de preparació possible. Tenint en compte la qualitat i contingut de les dades disponibles es decidirà quines poden ser rellevants i quines no. Un cop seleccionades les fonts d'informació importants es procedirà a netejar-les i preparar-les, és a dir, començar a transformar-les. Aquesta transformació la dividirem en dues fases diferents, neteja i integració.

4.3.1. Neteja

La neteja s'entén com el procés de tractar els valors incoherents o erronis presents en les dades. Aquests valors poden ser deguts a causes molt diverses, com per exemple, errors de lectura, humans o de format. Es poden dur a terme diferents transformacions en funció de les dades que es tinguin, eliminar valors erronis, uniformitzar valors absents, recerca de valors correctes, etcètera. Cada vegada que s'aplica una transformació aquesta ha de quedar documentada en un informe. Aquest informe, que es pot veure en la Taula 3, es realitzarà per a cada fonts d'informació i s'afegirà a l'inventari de dades, l'objectiu d'afegir-lo és tenir tota la informació sobre les dades en un mateix document.

4.3.2. Integració, millora i enriquiment

L'objectiu de la fase d'integració és obtenir tot el conjunt de dades netes centralitzades en una sola taula. La integració implica combinar dades provinents de diferents fonts per obtenir una visió unificada d'aquestes dades[10, 11]. Durant aquest procés és habitual trobar-se amb problemes d'heterogeneïtat semàntica [12], com ara, diferents indexacions entre les fonts, mateixos noms de variable que contenen diferents continguts o en diferents unitats, noms de variable diferents amb el mateix contingut, diferents tipus d'agregació entre fonts, diferents maneres d'expressar el mateix.

La millora de dades (*Data Enhancement*) i l'enriquiment de dades (*Data Enrichment*) [13], tenen com a objectiu aportar informació, addicional al conjunt de dades, que pugui ser d'utilitat. Tot i compartir el mateix objectiu aquests processos tenen diferents significats. S'anomena millora de dades quan s'expandeix el conjunt de dades sense fer ús de fonts

externes, per exemple aplicar ràtios o diferències entre variables de la taula. S'anomena enriquiment de dades quan s'afegeixen noves variables provinents de fonts externes, que tenen algun tipus de relació amb alguna de les variables de la taula, per exemple, afegir dades meteorològiques a partir de la zona geogràfica.

Un cop s'ha integrat, millorat i enriquit, l'analista disposarà de la *Taula base d'anàlisi*, en ella s'hi emmagatzemen totes les dades que posteriorment s'analitzaran.

4.4. Anàlisi preliminar

L'objectiu és extreure coneixements subjacents a les dades, que puguin ser d'utilitat. En aquesta fase s'apliquen, des de simples anàlisis estadístics a models analítics més elaborats, sobre les dades contingudes en la taula base d'anàlisi. Amb l'aplicació d'aquests models es busquen correlacions, patrons, anomalies o valors que puguin ser d'interès. Els resultats extrets a partir de l'anàlisi han de servir per poder valorar la viabilitat d'implementar models analítics més elaborats. Cal tenir en compte que es possible trobar-se amb conclusions que puguin ser obvies, però com a norma general aquestes conclusions no es podien extreure o corroborar anteriorment, ja que les dades estaven disperses i era impossible realitzar-hi cap tipus d'anàlisi.

És possible que per aplicar diversos models sigui necessària una nova preparació per poder extreure la màxima informació possible, és per això, que en la Figura 3, en que es mostra el diagrama de la metodologia, es pot observar com aquesta fase pot retroalimentar l'anterior fase en que es preparen les dades.

4.5. Avaluació, recomanacions i desplegament

L'objectiu és plasmar els coneixements extrets durant el procés, proposar un pla de millora, i crear unes directrius per seguir l'anàlisi.

Es crea un document on es plasmen tots els coneixements extrets de l'anàlisi així com recomanacions necessàries per millorar la gestió de la informació. També es crea un full de ruta amb les següents accions a realitzar per tal de seguir fent DM en la direcció més adequada.

5. Comprensió del negoci

Per la realització d'aquesta fase i seguint la metodologia proposada, primer s'ha procedit a la lectura de la documentació existent, acte seguit s'han realitzat entrevistes amb treballadors de l'empresa, seguint un ordre de menor a major responsabilitat. En la Figura 6 es poden veure les accions que s'han dut a terme.

5.1. Introducció a l'empresa Enertika

Enertika, és una empresa de serveis energètics del tipus *ESCO (Energy Service Companies)* [14, 15]. Aquest tipus d'empreses tenen uns trets característics que les diferencien de les consultores energètiques tradicionals.

- Garanteixen una millora en la eficiència, i/o una provisió del mateix nivell d'energia a un menor cost, de manera que els estalvis obtinguts siguin suficients per afrontar els costos dels projectes.

- La seva remuneració està directament lligada als nivells d'estalvi energètic obtinguts.

- Majoritàriament financen o ajuden a obtenir finançament per a l'execució dels projectes.

En aquest tipus d'empresa els projectes consten d'un procés d'estudi, d'execució i d'operació i manteniment. Aquest últim és molt important, ja que és el que assegura l'obtenció dels estalvis energètics, i per tant, el retorn econòmic.

La majoria dels projectes duts a terme per Enertika han sigut en el sector de les telecomunicacions i en el sector públic. Un dels seu trets característics és la aposta pel monitoratge a temps real a través de Internet, seguint la filosofia *Internet of Things* (IoT). Les variables monitorades es posen a la disposició dels clients i s'usen internament per controlar tot el procés d'operació i manteniment dels projectes executats.

Tenint en compte els recursos disponibles per a la realització d'aquest treball, l'estudi es centrarà en el projecte Free-Cooling a Mèxic, i a partir d'aquest es valorarà la viabilitat d'estendre'l als altres projectes de l'empresa. L'elecció del projecte Free-Cooling ve argumentada per dues raons principals; la quantitat d'EB (estacions base), unes 1000 castes; i la data de la seva execució, l'any 2014, per tant es disposa d'una quantitat de dades important.

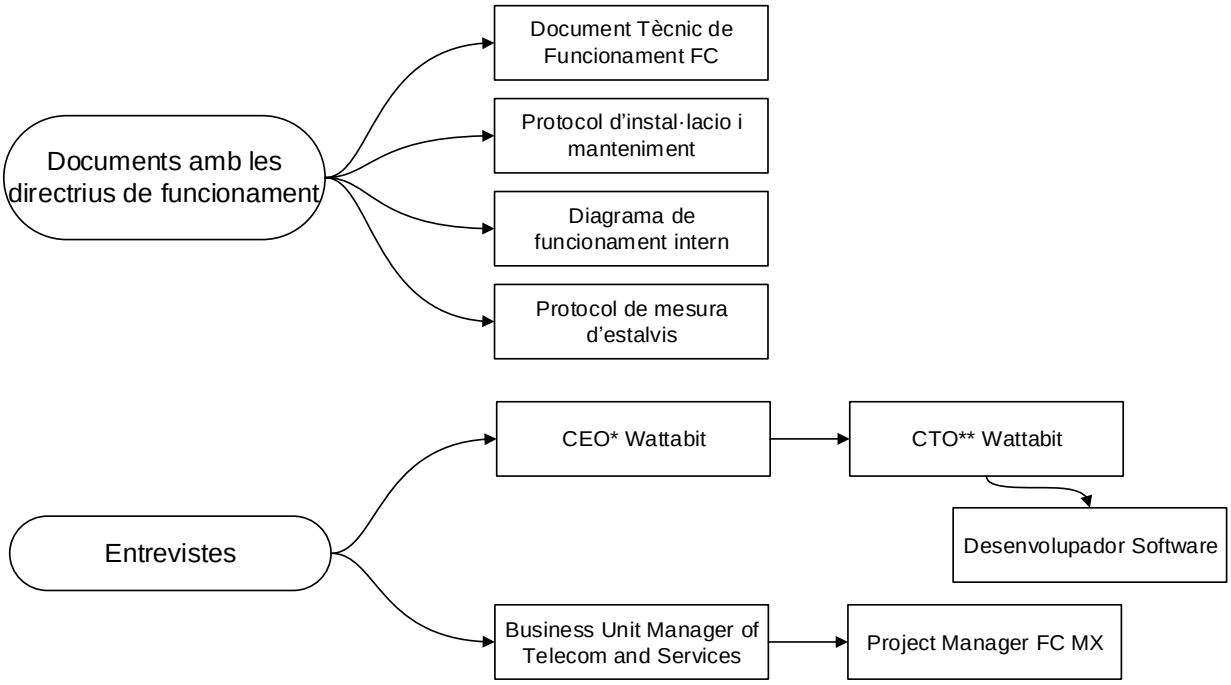


Figura 6. Accions realitzades en la fase de Comprensió del negoci

5.2. Introducció al projecte Free-Cooling

El projecte Free-Cooling [16] s'emmarca en el sector de les telecomunicacions, més concretament en la millora de l'eficiència energètica de la climatització de les Estacions Base de Telecomunicacions (EB).

Una EB és aquell lloc on hi ha els aparells necessàries per a mantenir les comunicacions telefòniques sense fil. Les EB poden ser *indoor* o *outdoor*, en aquest cas només es tindran en compte les *indoor*. Els aparells que contenen les EB es poden diferenciar en dos grup, aparells de transformació i emmagatzematge d'energia, i aparells de telecomunicacions pròpiament dits. Tots aquests aparells dissipen energia en forma de calor, de manera que és necessari gestionar l'eliminació d'aquesta calor. Tradicionalment aquesta problemàtica s'ha resolt amb la instal·lació d'aparells d'aire condicionat (AA), en alguns casos prioritzant la 'no caiguda' dels serveis, deguda a la temperatura dels aparells, en detriment dels costos energètic generats.

La solució proposada per reduir aquests costos energètics és la instal·lació d'un sistema de Free-Cooling (FC) que complementa i controla el sistema AA. Un sistema FC es podria definir com un sistema de ventilació natural forçada. El que fa aquest sistema és extreure la

calor mitjançant convecció forçada, és a dir, fer circular un flux d'aire a menor temperatura (de l'exterior) en l'interior de la EB. Per tant aquest mecanisme d'extracció de calor només és funcional quan la temperatura exterior és menor a la interior, quan aquesta circumstància no es dóna, o quan el FC no és capaç d'extreure suficient calor, es fa ús del sistema AA.

Complementàriament s'instal·la un sistema capaç de monitorar totes les variables de funcionament del FC i el consum energètic de l'EB. Aquestes dades s'envien cada quart d'hora a la plataforma de monitoratge Wattabit, per tal de poder fer un seguiment a temps real del funcionament de les EB i el sistema FC i poder detectar-hi incidències.

La solució implementada consta de diversos aparells:

- Ventiladors: Extreuen aire de l'interior.
- Comportes de sobrepressió: Deixen entrar aire de l'exterior quan els ventiladors estan en funcionament.
- Sondes tèrmiques (NTC): Mesuren la temperatura interior i exterior del sistema.
- Contactors: Controlen l'alimentació elèctrica de l'aparell AA
- Comptador d'energia: Mesura les dades de consum energètic de la EB. Va acompanyat dels corresponents transformadors d'intensitat necessaris.
- PLC: Element electrònic que controla el funcionament de tot el sistema seguint una lògica de funcionament.
- Mòdem: Transmet les dades entre el PLC i la plataforma de monitoratge.

5.3. Situació i funcionament actuals

Actualment el projecte FC Mèxic està en una etapa d'operació i manteniment (O&M), per tant ja s'han instal·lat la totalitat de les EB de l'abast del projecte. En aquesta etapa es fa un seguiment per tal de comprovar, mantenir i millorar el bon funcionament de la solució; i informar l'empresa propietària de les instal·lacions. El diagrama de funcionament d'aquesta etapa es pot veure en la Figura 7.

La valoració del bon funcionament es fa bàsicament a partir de tres fonts d'informació diferents:

- Alarmes WTB: Aquestes alarmes s'han creat de manera que es rebin notificacions en forma de alarma quan el funcionament no és l'adequat. Per exemple: quan el

consum o temperatura sobrepassen un valor llindar, quan es perd la comunicació amb algun dels centres, quan hi ha errors en les sondes de temperatura o en el comptador d'electricitat...

- Desviacions entre CFE i WTB: La desviació entre les lectures de la companyia subministradora d'electricitat i les lectures del comptador instal·lat, del qual se'n fa el monitoratge.

- Manteniments Preventius: Quan es detecta un mal funcionament de manera presencial al realitzar el manteniment preventiu.

Aquest seguiment del projecte provoca bàsicament dues accions presencials, aquestes accions són:

- Manteniments Correctius: Aquestes accions es duen a terme quan s'observa un comportament anòmal de les variables monitorades. Aquestes variacions en el comportament poden ser d'índole molt diversa, des d'interrupcions en el subministrament energètic, fins a manipulacions del contactor o PLC encarregats del funcionament del FC.

- Manteniments Preventius: Aquestes accions estan planificades de manera que es visitin totes les EB de manera periòdica, per així obtenir coneixement de l'estat de les EB i els components del FC. Un dels objectius principals d'aquests manteniments és evitar que apareixien possibles avaries que comportin un manteniment correctiu.

Després de dur a terme cada manteniment es crea un informe on s'explica què s'ha fet en la intervenció. Per tal de tenir un control sobre aquests manteniments es fan servir uns codis únics anomenats Ordres de Treball (OT), aquests codis s'usen per tenir una millor traçabilitat dels manteniments que es sol·liciten i es realitzen.

Per altra banda també es realitzen Informes d'estalvis periòdics, trimestrals i anuals, en aquests informes s'hi recull tota la informació que sigui rellevant per l'empresa propietària de les EB. L'objectiu d'aquests informes és mostrar al client el valor afegit aportat per Enertika, i l'estalvi generat, així com qualsevol incidència que s'hagi pogut detectar.

L'estalvi es mesura seguint les directrius del IPMVP [17] (*International Performance Measurement & Verification Protocol*) i l'acord [18] arribat amb l'empresa propietària de les EB. Per extreure l'energia estalviada es compara l'energia actual amb l'energia registrada per CFE en aquell mateix mes, quan el FC encara no s'havia implementat, aquest consum s'anomena línia base (LB). Aquest càlcul es mostra en la següent equació.

$$Estalvi\ \% = \frac{Línia\ Base_{mes} - Consum_{\ mes}}{Línia\ Base_{mes}}$$

En veure que s'instal·larien nous equips LTE a la gran majoria de EB es va acordar amb el client que aquest sobre consum, degut al LTE, es restaria al consum total en compte de ser sumat a la línia base.

$$Estalvi\ \% = \frac{Línia\ Base_{mes} - (Consum\ Total_{\ mes} - Consum\ LTE_{mes})}{Línia\ Base_{mes}}$$

Es calcula d'aquesta manera, ja que sinó repercutiria negativament en el càlcul dels percentatges, i els objectius del projecte estan estipulats de manera percentual, l'estalvi objectiu del projecte és del 25%.

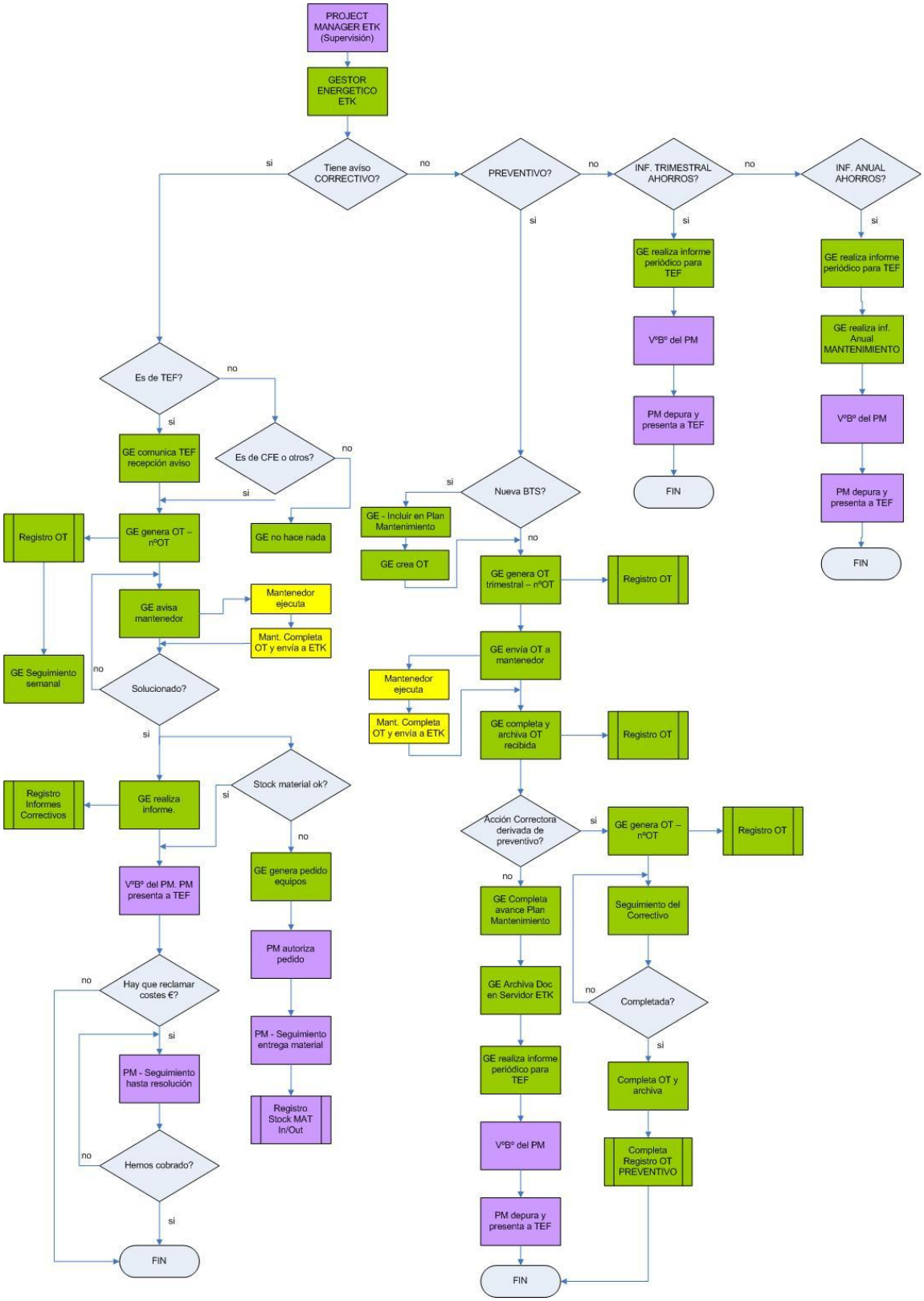


Figura 7. Diagrama de flux de treball de l'etapa d'operació i manteniment

6. Auditoria i comprensió de dades

6.1. Exploració de dades

6.1.1. Repositoris de dades

L'empresa fa ús de tres repositoris diferents, cadascun amb diferents usos. Aquests repositoris són els següents:

- Servidor intern d'expedients: En aquest servidor s'hi haurien de guardar tots els documents referents a cada expedients o projectes. L'objectiu d'aquest servidor és emmagatzemar totes les dades generades durant el disseny, execució, operació i manteniment dels projectes.

- Servidor Google Drive: En aquest servidor s'hi emmagatzema informació operativa i s'usa com a via de comunicació entre Barcelona i Mèxic.

- Base de dades Wattabit: En aquesta base de dades s'hi carreguen totes les dades provinents dels data loggers. Aquests data loggers emmagatzemen els valors de totes variables de funcionament de la EB i les envien cada 15 minuts a través de connexió telefònica. L'objectiu d'aquest repositori és contenir els valors de les variables de funcionament de les EB en relació al temps.

6.1.2. Informació existent

En la Figura 8 s'hi representa la informació que es genera i s'ha generat durant el transcurs de les diverses etapes del projecte, seguint la notació proposada en l'apartat 4.2.1. La informació existent mostrada en la Figura 8 es pot dividir en dos tipus de provinences:

- Informació externa. Proporcionada per la Comissió Federal d'Electricitat i per Telefònica.

- Informació interna. Prové dels sensors instal·lats en les EB, els algoritmes de tractament de la informació dels sensors de Wattabit i per últim tota la informació generada per Enertika durant l'execució i l'operació i manteniment. Es representa en el diagrama de color verd.

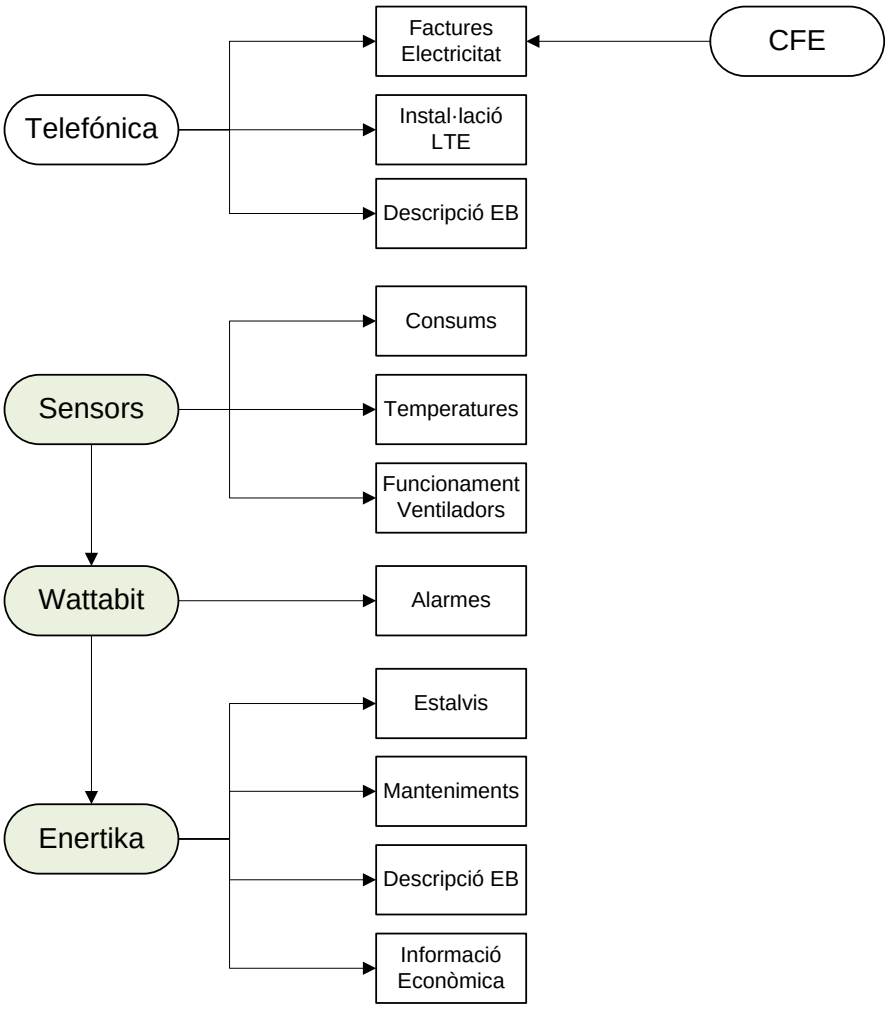


Figura 8. Diagrama d'informació existent

6.1.3. Flux d'informació operativa

En la Figura 9 s'hi representa el flux d'informació entre Enertika, la Comissió Federal d'Electricitat i Telefònica. Tot i tenir aquests fluxos definits i amb una freqüència determinada, hi ha hagut interrupcions en els fluxos. És important tenir aquests fets presents, ja que poden influir en la qualitat i/o la coherència de certes dades. Cal destacar dos tipus d'interrupció, provinents de Telefónica.

- Informació dels canvis en els equips instal·lats en les EB. S'ha deixat de rebre informació real i contrastada per rebre estimacions i finalment no rebre cap tipus d'informació.

- Factures d'electricitat de totes les EB. La freqüència d'enviament no s'ha respectat, dificultant el correcte seguiment.

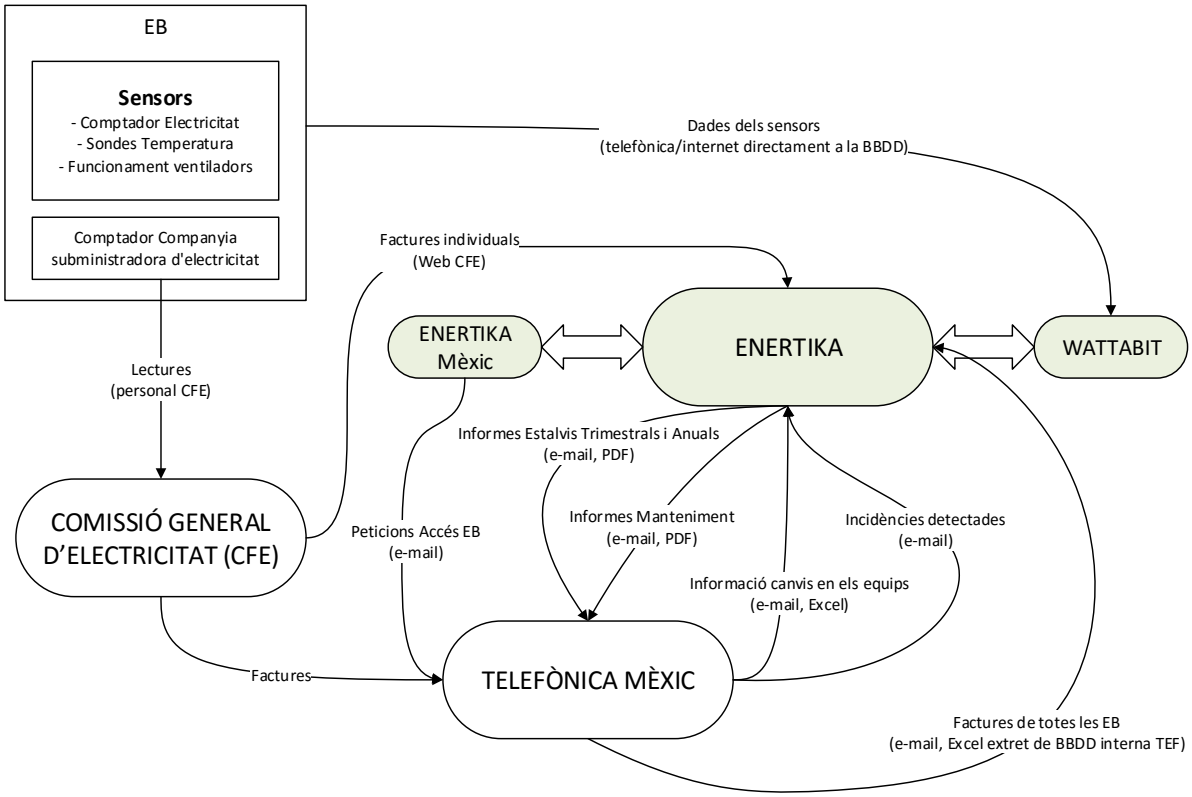


Figura 9. Flux d'informació operativa

6.2. Inventari de dades

Per tal d'auditar i comprendre les dades disponibles, seguint la metodologia establerta, s'ha creat el document *Inventari de Dades* per a cadascuna de les fonts d'informació. El primer apartat corresponent a les *Metadades* s'ha omplert a partir de les característiques de l'arxiu i els coneixements dels treballadors que fan ús o són responsables dels arxius, se'n pot veure un exemple en la Taula 4. Posteriorment s'ha omplert la informació requerida per l'*Estudi de dades*. En la Taula 5 es pot observar un exemple d'*Estudi de dades*.

Tenint en compte que la gran majoria de dades estaven emmagatzemades en format Excel s'ha creat un *script* de Python que extreu les dades necessàries per omplir l'*Estudi de dades* a partir dels documents d'Excel, de manera automàtica. Aquest *script* crea un arxiu de text i un histograma de freqüència per a cadascuna de les variables de l'Excel, aquests es poden observar en la Figura 10. En l'arxiu s'hi emmagatzema el nombre d'entrades; la mitjana; la desviació estàndard; els valors mínims, màxims i dels quartils; el nombre de valors únics, i els seus valors (en cas de ser pocs). Si s'executa en una variable no numèrica retorna la mateixa informació referida a la freqüència.

```
Describe_____________________________________________

count    1039.000000
mean        7.692974
std         1.525335
min         3.000000
25%         7.000000
50%         8.000000
75%         9.000000
max         9.000000
Name: Region, dtype: float64


Uniques______________________________________________

Uniques = 4

9    407
7    381
8    182
3     69
Name: Region, dtype: int64
```

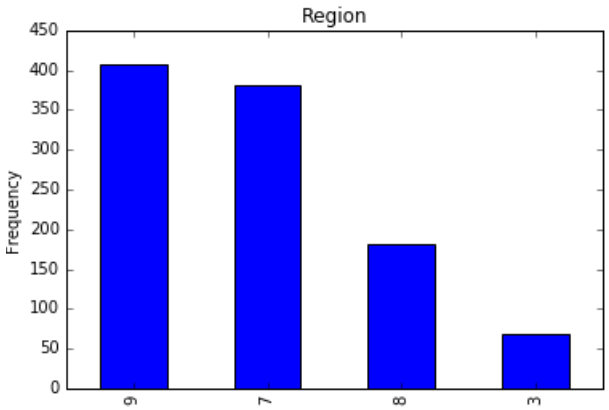


Figura 10. Exemple de l'arxiu de text i l'histograma de freqüència generats a partir de la variable Regió

METADADES	ID	00002	Nom	Previsió LTE	Data	26/03/2016
Descripció	Informació sobre les dates previstes per Telefónica México per instal·lar equips LTE en 2279 centres (només una part d'ells forma part del projecte FC MX)					
Creador	Telefónica México					
Font	Previsió interna de Telefónica México en quant a la instal·lació de LTE					
Responsable/Gestor	Project Manager (Enertika)					
Funció	Conèixer la previsió d'instal·lació de l'empresa per poder comprovar si hi ha augments en el consum					
Format	Excel					
Ubicació	Servidor Enertika (ETK\70031- TEF MX FC O&M\2 Datos técnicos proyecto\2.7 Gestión Energética\2.7.7 Estudio 4G\ENERTIKA fechas LTE.xls)					
Restriccions	Hi pot accedir tot el personal d'Enertika que tingui accés complet al servidor d'expedients					
Actualitzacions	No s'han obtingut noves actualitzacions de les previsions de Telefònica tot i que s'haurien hagut de rebre					
Rèpliques i Còpies	Còpia del document en el Servidor Goolge Drive (Drive\FC MX\Seguimiento\ESTUDIO 4G\ENERTIKA fechas LTE.xls)					
Fiabilitat	Les dates d'instal·lació no són fiables degut a que són previsions, però aporten una visió aproximada					
Comentaris	Hi ha informació de 2279 centres, caldrà comprovar quants d'aquests formen part del projecte FC MX					

Taula 4. Metadades de la font d'informació Previsió LTE

ESTUDI DADES					ID	00002		Nom	Previsió LTE		Data	26/03/2016
Variables	**Col.**	**Mesura**	**Escala**	**Descripció**	**nº**	**Únic**	**Tipus**	**Max**	**Min**	**Mean**	**StDev**	**Observacions**
Regió	1	-	Categòrica	?	2279	8	enter	9	1	6,5304	2,6221	
ID	2	-	Nominal	Codi numèric d'identificació de la EB	2279	2279	text	1	1	1	0	
Nom	3	-	Nominal	Nom d'identificació de la EB	2279	2091	text	11	1	1,0899	0,4348	
Mercat LTE	4	-	Categòrica	Regió de mercat (classificació Telefónica)	2279	19	text	891	24	119,95	195,09	
Tecnologia	5	-	Categòrica	Tecnologia que s'ha d'instal·lar (4G [LTE] en tots els casos)	2183	1	text	2183	2183	2183	0	No aporta informació útil
Plan	6	Any	Temporal	Any en que està previst el pla d'execució	2279	2	enter	2015	2014	2014,1	0,3289	
Data	7	DD/MM/AAAA	Temporal	Data prevista per a la instal·lació	2279	294		46	1	7,7517	7,4867	La freqüència segueix una distribució quasi exponencial

Taula 5. Estudi de dades de la font d'informació Previsió LTE

6.3. Resum de la situació

6.3.1. Informació extreta

Informació global sobre la tipologia de les EB

S'ha pogut observar una variabilitat relativament baixa en les qualitats de les EB, per tant es pot parlar d'una tipologia estàndard dins d'aquest projecte. Com es pot veure en la Taula 6 i en la Figura 11, la gran majora de EB són de 7 m^2, amb bateries d'entre 500 i 600 Ah i fetes amb *multipanel*.

Àrea (m^2)	Bateries (Ah)	FORMIGÓ	MULTIPANEL
7	<500	0,00%	0,11%
	500-600	10,53%	67,02%
	>600	0,53%	0,43%
8	600	0,00%	0,11%
8,9	<500	0,00%	0,11%
	500-600	0,00%	13,94%
9	585	0,11%	0,00%
10	500-600	1,81%	4,68%
	>600	0,53%	0,00%
12	600	0,11%	0,00%

Taula 6. Distribució de les tipologies de les EB (940)

Variables no documentades

En alguns casos s'ha topat amb variables de les quals se'n desconeixia el significat o aquest no era obvi. Un dels casos és la variable Regió, aquesta està representada per quatre valors enters diferents, provinents d'una classificació realitzada per TEF. Com es pot veure en la Figura 11, aquesta variable discretitza la localització de les EB en quatre zones. A priori sembla que els valors enters no tenen cap relació, amb la localització, ni longitudinalment ni transversalment.

Un altre cas de variable no documentada és la variable Zona Climàtica, aquesta variable discretitza la variable Població en 19 grups diferents. Aquests grups són diferents ciutats de les quals es té informació meteorològica històrica.

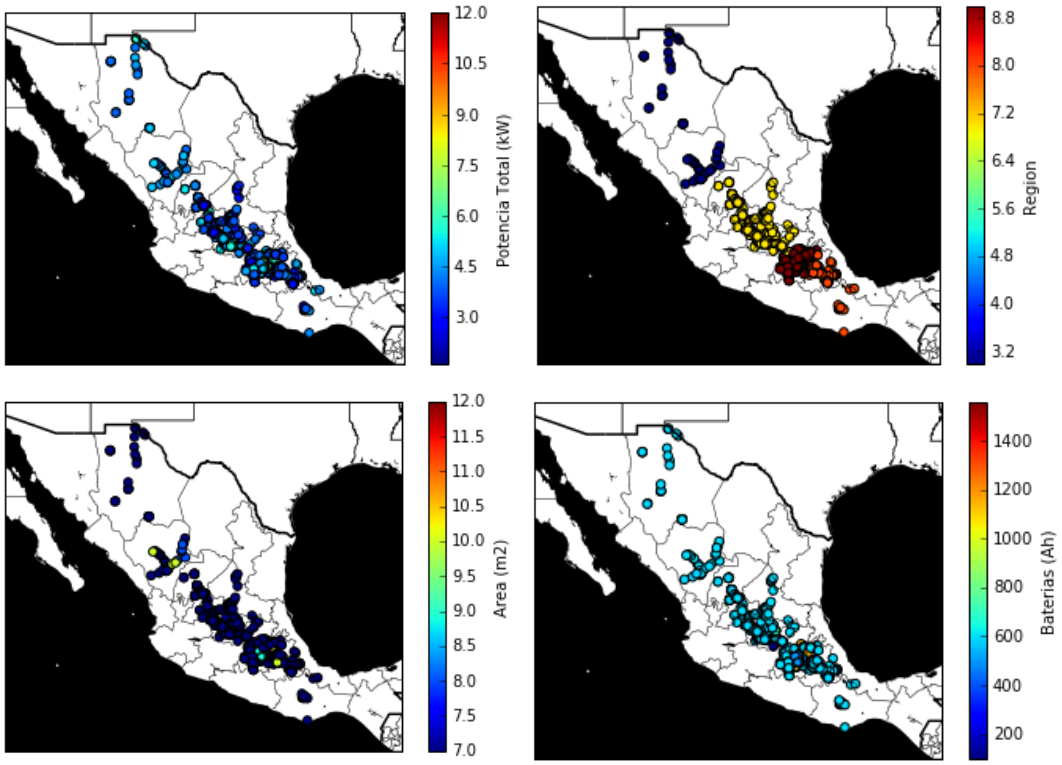


Figura 11. Representació en el context geogràfic de les variables Potència Total, Regió, Àrea i Capacitat de les bateries

6.3.2. Problemàtiques

Organitzativa

- Ús de diferents repositoris, tot i que en un principi s'usin per fins diferents fins i no hagin d'interferir, és fàcil que es creïn duplicats i versions desactualitzades dels documents.

- Gestió incorrecta de la jerarquia de carpetes. Això repercuteix negativament no només en la cerca de fitxers sinó també en el seu correcte desat. Això pot produir la duplicació de continguts.

- Manca d'estandardització a l'hora d'anomenar versions actualitzades i a l'hora de desar versions desactualitzades. A dia d'avui s'usa una gran varietat de marques indicatives de la versió, arribant fins i tot a solapar-se. Alguns exemples d'aquests marques són: *any mes dia* (nom), *dia mes any* (nom) , (nom) - *any mes*, *setmana any* (nom), (nom) rev *X*, etcètera, tot i que en alguns casos també es fa ús d'una carpeta, anomenada 'old', on s'emmagatzemen les versions antigues.

Obtenció d'informació, qualitat i periodicitat

- Base de dades de WTB, en aquesta base de dades s'emmagatzema tota la informació provinent dels sensors de la EB. Actualment no hi ha una eina d'extracció per obtenir totes les dades d'un projecte determinat. Es va fer arribar una petició al departament IT (*Information Technology*) per desenvolupar una eina per extreure les dades directament de la BBDD, és a dir, com a *raw data* o dades sense tractar. La petició va ser rebuda de manera positiva, però malauradament va ser impossible d'acomplir per culpa de la quantitat de tasques amb major prioritat assignades a aquests departament.

- Alarmes de WTB, la concepció de les alarmes va ser purament operacional i estan implementades com a notificacions, de tal manera que no estan emmagatzemades en la BBDD on hi ha les dades de cada EB. Aquest fet en dificulta la extracció i integració. Cal destacar que al tenir una funció operacional van ser eliminades en alguns períodes, de manera que no hi ha un històric fiable.

- Interrupcions en el seguiment dels protocols de registre d'informació dels registres de manteniments. Això provoca que hi hagi períodes sense informació. Aquesta alta dispersió pot provocar que les dades estiguin altament esbiaixades.

- Incompliments en l'enviament d'informació de les factures de CFE que TEF envia des de la seva base de dades. En alguns casos no es compleix el període d'actualització trimestral.

6.3.3. Solucions i recomanacions

Organitzativa

- Implementar una cultura entre els treballadors per potenciar la precaució i consciència que cal tenir quan s'usen dos repositoris. Assegurar-se que es controlen els arxius creats o modificats, de manera recurrent i es guarden al servidor intern. Aquesta opció suposa un esforç baix.

- Facilitar l'accés al servidor des de l'exterior de l'empresa via Internet. Una de les solucions es crear una VPN (*Virtual Private Network*) per poder accedir a la xarxa local a partir de la xarxa pública. És vital que es creï amb el màxim rigor per no comprometre la seguretat del servidor. La creació i manteniment de la VPN té un cost associat tot i que aquest no es considera elevat.

- Transmetre als treballadors la importància de seguir els protocols d'emmagatzematge d'informació establerts. Aquests protocols han d'estipular el funcionament de la jerarquia de carpetes, la manera estàndard d'anomenar els documents. Paral·lelament s'hauria de valorar la usabilitat dels protocols actuals per millorar-los i per desambigüitzar-los si es dóna el cas.

Obtenció d'informació, qualitat i periodicitat

- Crear una eina d'extracció massiva de dades directament des de la BBDD (solució a curt termini). La implementació d'aquesta solució suposa una inversió en recursos bastant continguda, s'estima que es podria acomplir focalitzant els esforços de treball del departament de IT sense invertir una gran quantitat de temps. Per altra banda, el beneficis aportats a altres departaments serien instantanis, disminuint el temps necessari per realitzar diverses tasques recurrents.

- Una solució més ambiciosa seria canviar el model actual de BBDD per apropar-lo cap al de DW. En aquest cas seria interessant integrar-hi altres dades, com per exemple les alarmes generades, d'aquesta manera deixaria de ser informació volàtil (solució a mig-llarg termini). La implementació d'aquesta solució suposa un esforç econòmic i temporal elevat, ja que un projecte d'aquestes característiques requereix d'un estudi en profunditat i personal especialitzat en la matèria. En cas de dur a terme un projecte d'aquestes magnituds és molt important que es destinin els recursos necessaris per a realitzar un bon estudi previ, ja que qualsevol canvi a posteriori suposa una quantitat molt més elevada de recursos.

- Estudi dels protocols establerts pel registre d'informació, per tal d'avaluar si són adequats, i donar resposta a perquè s'han deixat de respectar durant certs períodes. Depenent de la situació trobada pot acabar amb una simplificació dels protocols actuals per adequar-los a la realitat o transmetre als treballadors la importància de seguir el protocols establerts per registrat les actuacions. L'esforç per dur a terme aquesta acció és baix, però cal que hi estiguin implicats tant els treballadors com la direcció.

- Transmetre al client o empresa col·laboradora la importància de l'enviament periòdic de dades acordades per poder aportar valor afegit. Revisar quins són els punts acordats en el contracte, per conèixer si aquest intercanvi d'informació hi consta o no. L'esforç per realitzar aquesta acció es pot considerar nul.

6.3.4. Conclusions i pròxims passos

De totes les problemàtiques trobades la que es valora de manera més negativa és la relacionada amb la dificultat d'extracció de dades en brut emmagatzemades en la base de dades, ja que això suposa que no es pot obtenir més informació de les dades que la proporcionada per la plataforma Wattabit. Aquest fet minva de manera molt important el potencial de tenir un històric de dades d'aquestes dimensions. Degut a la importància de la informació continguda en aquestes dades i la única possibilitat d'obtenir-la mitjançant mètodes extremadament lents, s'ha decidit obtenir-la amb una agregació mensual. L'obtenció amb aquest nivell d'agregació és el factor que la fa viable des del punt de vista temporal. Cal tenir en compte que al tenir les dades agregades la seva fiabilitat pot disminuir.

Per prosseguir amb la integració i posterior anàlisi es creu convenient prescindir de les següents fonts d'informació:

- Descripció EB (provinent de ETK): Aquesta informació prové de l'execució del projecte, i està emmagatzemada en forma de informes (*Checklists*) en format PDF i fotografies. Tot i que gran part de la informació ja està continguda en la "Descripció EB" de TEF, conté informació que podria ser útil, com per exemple la disposició dels elements FC en l'interior de la EB. Es desestima l'ús d'aquesta font d'informació degut la gran inversió de temps que requeriria envers el valor que aquesta podria aportar. Però cal tenir-la present de cara poder ratificar hipòtesis futures.

- Alarmes WTB: Aquesta informació es desestima bàsicament per la elevada dificultat a l'hora d'extreure-la i pels períodes incomplets que té. A la vegada també es vol deixar constància que els rangs de les alarmes es modifiquen en funció del temps per adequar-se al comportament de cada EB, un dels grans problemes és que aquestes modificacions no queden registrades enlloc.

- Manteniments: Aquesta informació es desestima per les interrupcions trobades en els registres de control de manteniments correctius i preventius. Per tant les dades que se'n podria extreure serien molt disperses i podrien estar altament esbiaixades.

Es procedirà amb les següents fonts d'informació:

- Descripció EB (provinent de TEF): Aquesta font aporta informació descriptiva de cada EB, aquesta pot ser útil per a categoritzar-les.

- Factures CFE: Aquesta font d'informació es considera útil, ja que mostra l'energia que la companyia subministradora està facturant per a cada EB.

- Dates d'instal·lació LTE: Aquesta informació es considera d'utilitat tot i que en molts

casos la informació sigui inexistent, és a dir, tot i tenir un alt nivell d'incompletesa.

- Estudis de sobreconsum degut a LTE: Aquesta informació es considera útil, ja que conté dates reals d'entrada en funcionament de nous equips LTE amb el seu consegüent increment de consum de la EB. Cal deixar constància que també conté un elevat nivell d'incompletesa.

- Dades provinents dels sensors (BBDD de WTB): En aquestes dades s'hi inclouen temperatures i consums registrats, degut a la problemàtica amb l'extracció de dades de la BBDD, es farà ús d'aquestes dades amb una agregació mensual.

- Línies Base: Aquesta informació conté el consum previ que es té en compte per calcular els estalvis generats pel FC.

7. Neteja, integració i preparació preliminar

7.1. Neteja

Per poder identificar cada EB de manera individual, és essencial assignar una variable com a identificador únic. En aquest cas l'única variable que no es repeteix en les EB i que apareix en totes les fonts d'informació, és el ID de Telefónica, per tant s'usa aquesta variable com a índex de la taula resultat d'aquest procés.

7.1.1. Descripció EB

Tenint en compte que aquesta informació no s'actualitza periòdicament, s'ha netejat i corregit les dades d'una manera menys automatitzada. Aquest tipus de neteja i correcció és molt laboriosa i consumeix una quantitat important de temps. En aquest cas s'han usat diversos mètodes de neteja de dades, aquests es comenten a continuació.

- **Uniformització de les cel·les absents**. Això ha sigut necessari, ja que es feien servir diferents indicadors de valor absent o *missing*, els indicadors existents eren els següents, 'no medida','Cell error Type', '//', '0', '-'. En aquestes cel·les s'hi ha assignat el valor *NaN*[2].

- **Supressió de cadenes de caràcters**. En algunes variables els valors numèrics anaven seguits de cadenes de caràcters que han sigut suprimides, per exemple en el cas de l'àrea, el valor anava precedit de 'M2' i 'M 2'.

- **Supressió i correcció de valors incoherents**. S'ha detectat una quantitat important de valors incoherents. No totes les incidències han pogut ser corregides. Per corregir els valors s'han usat diversos mètodes, com la comparació amb altres documents que conten la mateixa variable i la recerca a través d'Internet. Els dos exemples més representatius són:

    - S'han substituït les EB amb material '*Lamina*' per '*Multipane'*, ja que observant algunes de les EB es pot veure que la gran majoria estan fetes del mateix material.

[2] En totes les taules es representaran els *missing* amb valor *NaN*, de la llibreria Python *Numpy*.

- S'han modificat les coordenades, ja que en alguns casos enlloc de contenir la informació corresponent a les coordenades contenien l'estat de la EB, mentre que altres estaven situades en l'altre hemisferi degut a un error en el signe. Després de realitzar aquestes modificacions s'han creuat les localitzacions amb les desades en un altre fitxer.

- **Uniformització de delimitador decimal**.

7.1.2. Temperatura exterior i interior

En aquest cas la neteja és fàcil d'aplicar, ja que es té coneixement dels diferents valors enviats pel PLC en cas d'obtenir errors en la lectura de la sonda de temperatura. En funció del tipus de PLC instal·lat s'envien diferents valors, aquests valors són, '3276', '6552' i '−60'. L'únic problema que es presenta és que al tenir les dades agregades aquests valors d'error es poden difuminar. Per a realitzar la neteja s'ha decidit declarar com a valors no vàlids els no continguts en l'interval $[60\,,-10]$.

7.1.3. Consums registrats a WTB i línea base

En el cas dels consums, cal assignar l'ID a dues de les EB, ja que havia sigut substituït erròniament pel seu nom. Aquests ID s'obtenen creuant els noms de les EB i els seus ID. Per altra banda la gran quantitat de valors 0 fa pensar que no són realment un valor de consum nul, sinó un valor que ens indica que no s'ha llegit el consum, per tant s'eliminen tots els valors iguals a 0. També s'eliminen els valors inferiors a 150 kWh per la mateixa raó. També es detecta un valor de 139531 kWh, el qual correspondria a una potència de 188 kW, aquest valor es considera erroni i s'elimina.

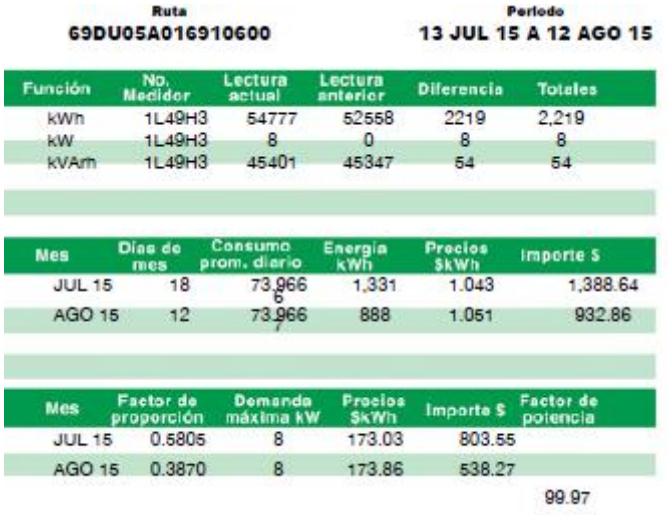
En el cas de les línies base quasi no es requereix neteja, ja que la LB és el valor de referència per a calcular l'estalvi, només cal eliminar algunes files que únicament contenen espais.

7.1.4. Consums CFE

En aquest cas durant l'inventari de dades es va veure que el nombre de dies no quadrava amb el nombre de dies continguts en el període de la factura. De manera que l'últim dia de la factura, el qual és el mateix que el primer de la següent factura es comptabilitza integrament les dues vegades. Com es pot observar en la Taula 7 entre la data d'inici i de fi de la factura (12/08/2015-13/07/2015) han passat 30 dies, mentre que si es sumen els dies assignats a cada mes (19+12) s'obtenen 31 dies. Al ser unes dades que s'obtenen regularment d'una BBDD de TEF no es tenia coneixement del tipus de tractament que aquestes rebien. Al disposar de l'accés a les factures individuals (Figura 12) es va realitzar

ID	Consum Diari	Consum Real	Any fact.	Mes fact.	Any Cobr.	Mes Cobr.	Dies	Inici	Final
24-03025	84,6667	1693,334	2015	7	2015	6	20	11/06/2015	13/07/2015
24-03025	84,6667	1100,6671	2015	7	2015	7	13	11/06/2015	13/07/2015
24-03025	71,5806	1360,0314	2015	8	2015	7	19	13/07/2015	12/08/2015
24-03025	71,5806	858,9672	2015	8	2015	8	12	13/07/2015	12/08/2015

Taula 7. Exemple de les dades de consum extretes de BBDD



Ruta
69DU05A016910600

Periodo
13 JUL 15 A 12 AGO 15

Función	No. Medidor	Lectura actual	Lectura anterior	Diferencia	Totales
kWh	1L49H3	54777	52558	2219	2,219
kW	1L49H3	8	0	8	8
kVArh	1L49H3	45401	45347	54	54

Mes	Días de mes	Consumo prom. diario	Energía kWh	Precios $kWh	Importe $
JUL 15	18	73.966	1,331	1.043	1,388.64
AGO 15	12	73.966	888	1.051	932.86

Mes	Factor de proporción	Demanda máxima kW	Precios $kWh	Importe $	Factor de potencia
JUL 15	0.5805	8	173.03	803.55	
AGO 15	0.3870	8	173.86	538.27	
					99.97

Figura 12. Exemple de factura de CFE

enginyeria inversa per deduir el tractament de dades. D'aquest procés es va comprovar que la suma dels consums reals coincidien (1360,0314+858,9672=1331+888) amb la factura per tant realment no es comptabilitzava el dia dues vegades de manera directa, sinó que s'usaven les següents equacions:

$$Consum\ Diari_{BBDD} = \frac{Comptador\ final - Comptador\ inicial}{Data\ Fi_{factura} - Data\ Inici_{factura} + 1}$$

$$Consum\ Real_{BBDD\ mes\ inici} = Consum\ Diari_{BBDD} * (Dies_{mes\ inici} + 1)$$

$$Consum\ Real_{BBDD\ mes\ fi} = Consum\ Diari_{BBDD} * Dies_{mes\ fi}$$

Cal fer constar que l'exemple mostrat es una particularització, ja que no totes les factures s'emeten mensualment, tot i això en els altres casos la problemàtica segueix sent la mateixa.

Per tal de revertir el tractament de dades fet per TEF es fa ús de les equacions:

$$Consum\ Diari = \frac{\sum Consum\ Real_{BBDD} \in Mateixa\ Factur}{Data\ Fi_{factura} - Data\ Inici_{factura}}$$

$$Consum\ Real_{mes\ inici} = Consum\ Diari * (Dies_{BBDD\ mes\ inici} - 1)$$

$$Consum\ Real_{mes\ inici} = Consum\ Diari * Dies_{BBDD\ mes\ fi}$$

$$Consum\ Mes\ Natural = \sum Consum\ Real\ \in Mateix\ Mes$$

Un cop realitzats aquests canvis s'ha procedit a obtenir una taula amb l'ID com a índex i les dates, mesos i anys, com a columnes. Alhora d'efectuar aquest canvi s'ha topat amb un problema, en certs casos particulars una EB disposava de més d'un RPU (número identificador del comptador elèctric de la companyia) i per tant de dos possibles consums pertanyents a un mateix mes i una mateixa EB. Aquest fet s'ha resolt sumant aquests consums.

7.1.5. Informació de la instal·lació 4G/LTE

En aquest cas no ha sigut necessari realitzar cap tipus de neteja, però si que ha sigut necessari un esforç important alhora de creuar dades entre diverses fonts i seleccionar els ID corresponents al projecte. Aquest aspecte es pot veure detallat en l'apartat 7.2.1.1.

7.2. Integració, millora i enriquiment

7.2.1. Integració i millora

Per comprovar la viabilitat d'integrar les dades de les diferents fonts es va començar creuant la informació entre totes les fonts, per poder tenir una idea clara del volum d'informació disponible associat a cada EB. Per dur a terme aquest creuament de dades es va fer ús de l'ID.

Inicialment es parteix dels ID que pertanyen a les EB executades en el projecte, i es comparen amb totes les taules. Com a primera exploració, si l'ID en qüestió apareix en la taula i conté algun tipus d'informació es comptabilitza obtenint la distribució percentual mostrada en la Taula 8. També es realitza un histograma que relaciona el nombre d'EB amb el nombre de taules on aquests apareixen, es pot veure en la Figura 13. Cal tenir en compte que d'aquesta manera no s'està avaluant el volum d'informació en la taula, sinó l'existència d'algun valor referit a l'ID en qüestió. D'aquesta exploració se n'extreu que el problema més

Taula	Percentatge
df_info	100,00%
df_CFE1	97,31%
df_Consum_WTB	89,61%
df_estalvi_noLTE	88,45%
df_LB	88,45%
df_temp_int	88,35%
df_temp_ext	87,97%
df_LTE_Prev	30,51%
df_consumLTE	28,01%
df_4G	11,26%

Taula 8. Distribució d'EB amb informació per taula

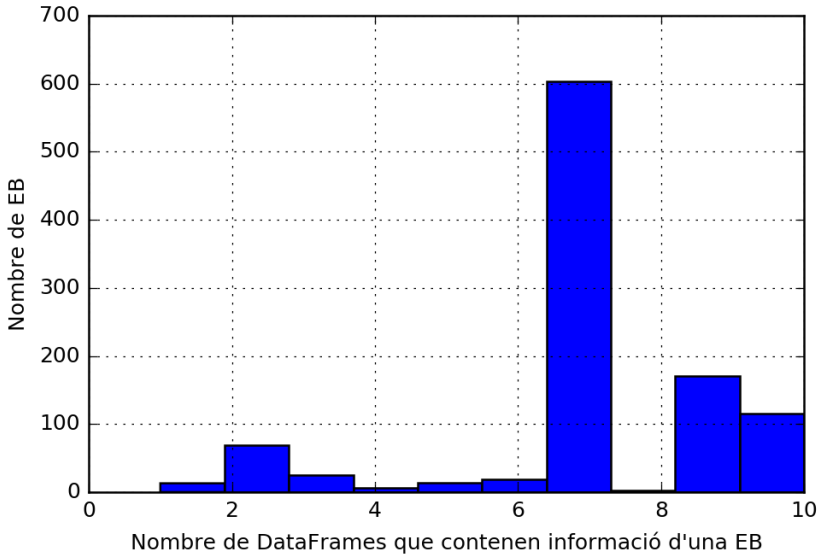


Figura 13. Distribució del nombre de EB envers el nombre de taules on apareix

greu es troba amb la informació que fa referència a la instal·lació d'equips LTE, ja que només es té informació de menys d'un terç de les EB del projecte. Per altra banda també es pot veure que 893 EB, és a dir, un 85,95% apareixen en 7 o més taules.

Totes les taules ja indexades per ID i que no requereixen de cap tipus de pivotatge o transformació, es concatenaran (segons la funció *concat* de la llibreria *Panda*) de la manera mostrada en la Figura 14, on la taula 1 sigui una taula que contingui indexats tots els ID de l'abast del projecte, en aquest cas es tracta del document que conté la descripció de cada EB. D'aquesta manera s'aconseguirà que en la taula definitiva no hi hagi cap EB que no formi part del projecte, però que formi part de les fonts de dades usades.

df1

	A	B	C	D
0	A0	B0	C0	D0
1	A1	B1	C1	D1
2	A2	B2	C2	D2
3	A3	B3	C3	D3

df4

	B	D	F
2	B2	D2	F2
3	B3	D3	F3
6	B6	D6	F6
7	B7	D7	F7

Result

	A	B	C	D	B	D	F
0	A0	B0	C0	D0	NaN	NaN	NaN
1	A1	B1	C1	D1	NaN	NaN	NaN
2	A2	B2	C2	D2	B2	D2	F2
3	A3	B3	C3	D3	B3	D3	F3

Figura 14. Exemple de concatenació de taules

7.2.1.1. Cas instal·lació 4G/LTE

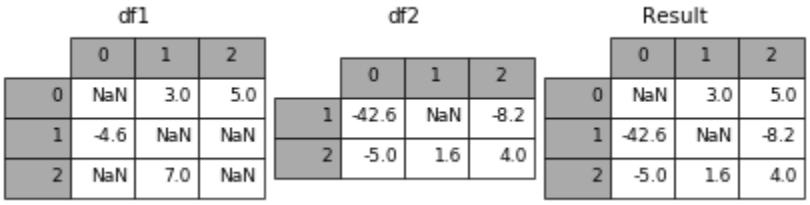
En aquest cas es disposa de diferents fonts amb informació sobre les mateixes variables, i en alguns casos amb diferents valors, per tant cal treballar amb aquesta informació d'una manera una mica més elaborada per tal de fusionar-ne el contingut.

Les variables a fusionar són la data d'entrada en funcionament dels equips LTE i l'increment de consum que van aportar. Les fonts d'informació obtingudes són tres, una d'elles conté les dates d'instal·lació estimades durant la planificació que va realitzar TEF, la segona conté un estudi realitzat a 118 EB a través de les eines de monitoratge, i per últim es té un segon estudi realitzat a un major nombre de EB i amb un nivell de precisió menor i que pot contenir EB del primer estudi. Per tant la fiabilitat de les fonts, classificada de major a menor, queda de la següent manera, *Estudi 4G (118)*, *Estudi 4G*, *Previsió TEF*.

Per tal de combinar la informació es va fer ús de la funció *Update*, aquesta funció actualitza els valors d'una taula a partir d'una segona taula, sempre respectant l'indexat, que ha de ser únic, i les columnes de la primera taula. El seu funcionament es pot veure en la Figura 15.

En primera instància es va crear una taula indexada amb tots els ID de l'abast del projecte i amb dues columnes. Una referent a la data de posada en funcionament, plena d'elements temporals nuls *NaT*[3], i la segona referent a l'increment de consum plena d'elements nuls *NaN*. Per aconseguir la taula desitjada es va anar actualitzant la taula buida amb les altres taules seguint un ordre de menor a major fiabilitat, per poder obtenir el major nombre d'elements fiables possibles. En la Taula 9 es pot veure l'evolució del nombre de valors i valors únics després de cada actualització. Cal remarcar que tots els valors d'increment tenen una data associada, per tant es té informació d'incremet de consum i data d'instal·lació per a un 28,1% de les EB del projecte.

[3] El valor *NaT* és l'equivalen al valor *NaN* per a objectes *datetime* de la llibreria Python amb el mateix nom.



	0	1	2
0	NaN	3.0	5.0
1	-4.6	NaN	NaN
2	NaN	7.0	NaN

	0	1	2
1	-42.6	NaN	-8.2
2	-5.0	1.6	4.0

	0	1	2
0	NaN	3.0	5.0
1	-42.6	NaN	-8.2
2	-5.0	1.6	4.0

Figura 15. Exemple d'actualització de taules

Actualització		1	2	3
Data_LTE	Valors	317	320	320
	Únics	123	141	141
Consum_LTE_dia	Valors	0	291	292
	Únics	0	86	106

Taula 9. Evolució del nombre de valors i valors únics després de cada actualització

Un cop aconseguida la taula desitjada es va procedir a millorar les dades de data d'instal·lació. Per fer això es va crear una columna per a cada mes, de la mateixa manera que estan emmagatzemats els consums mensuals, en cada cel·la s'hi calcula un valor del 0 al 1. On aquest valor és la ràtio de dies amb instal·lació LTE envers els dies del mes, per tant 0 indica que encara no s'han instal·lat equips, 1 que han estat instal·lats durant tot el mes.

Posteriorment es va crear una taula amb els sobreconsums mensuals associats als equips LTE fent ús de l'equació:

$$Consum_{LTE} = Ràtio\ Funcionament_{LTE} * Increment\ Consum_{Diari} * Dies_{mes}$$

Un cop obtinguda aquesta taula es va comprovar si alguna de les cel·les amb contingut *NaN* tenia algun valor associat en la taula obtinguda de l'Excel que contenia els valors de sobreconsum per mes i EB. El resultat d'aquesta comprovació va ser l'esperat, és a dir, cap dels valors *NaN* tenia cap valor associat, gràcies al resultat d'aquesta comprovació no cal realitzar cap tipus de combinació entre les taules, ja que la taula recalculada es considera més fiable i extensa al ser extreta de diferents fonts, tal i com anteriorment s'ha explicat.

7.2.2. Enriquiment

L'enriquiment és l'acció d'afegir noves variables, provinents de fonts externes, en la taula base d'anàlisi. Usualment, s'afegeixen variables sospitoses de tenir alguna relació o influència amb alguna de les variables no enriquides. L'objectiu de l'enriquiment és aportar informació que pugui ser d'utilitat en el moment de realitzar l'anàlisi.

- **Altitud**, per obtenir aquesta variable es fa ús d'una base de dades geogràfica del tipus DEM, *Digital Elevation Model,* més concretament de la base de dades NED1, *National Elevation Dataset,* proporcionada pel USGS, *United States Geological Survey*. Aquesta base de dades conté la informació referida a l'altitud amb una resolució de 30 metres. Aquesta informació s'extreu amb l'ajuda d'un aplicació web de la pàgina web *GPSVisualizer* [19]. Aquesta informació també es podria extreure d'una manera més automatitzada fent ús d'alguna API, com per exemple la de *Google Maps*.

- **Conductivitat tèrmica**, aquesta variable s'obté a partir del material assignat a la EB. Es tenen en compte dos materials. El primer és l'anomenat Multipanel, aquest material no és mes que dues fines plaques d'alumini, acer galvanitzat o fibra amb una capa interna d'escuma de poliuretà. Seguint les consideracions preses en l'estudi '*Heat flows and energetic behavior of a telecommunication radio base station'* [20], només es té en compte la conductivitat tèrmica del poliuretà, i es pren el valor 0,021 W/m^{2}K. El segon material és Formigó, en la gran majoria dels casos es tracta de parets prefabricades de formigó, per assignar la seva conductivitat tèrmica es té en compte la ponència '*Thermal and energetic analysis of a precast panel for industrial building'* [21] on es determina una conductivitat de 0,85 W/m^{2}K pels blocs de formigó prefabricats que s'usen per a parets.

- **Ràtio d'hores de sol**, aquesta variable s'obté de manera matemàtica partir de la latitud de cada EB. Les següents equacions determinen les hores de sol depenent del dia de l'any i la latitud de la localització.

$$x = \begin{cases} 0.8333, & Centre\ del\ sol\ en\ l'horitzó \\ 0.2667, & Extrem\ del\ sol\ en\ l'horitzó \\ 0.0\quad , & Extrem\ del\ sol\ aparentment\ en\ l'horitzó \end{cases}$$

$$P = \sin^{-1}\left(0.39795\cos\left(0.2163108 + 2\tan^{-1}\left(0.9671396\tan\left(0.0086(Dia - 186)\right)\right)\right)\right)$$

$$Hores\ Sol = 24 - \frac{24}{\pi}\cos^{-1}\left(\frac{\sin\left(\frac{x\,\pi}{180}\right) + \sin\left(\frac{Latitud\,\pi}{180}\right)\sin(P)}{\cos\left(\frac{Latitud\,\pi}{180}\right)\cos(P)}\right)$$

Ciutat de referència	Temp. Max	Temp. Mitja	Dies sup. 30ºC	Dies sup. 20ºC
San Luis Potosí	37,00 ºC	18,06 ºC	60	335
Veracruz	44,00 ºC	25,92 ºC	189	360
Puebla	31,00 ºC	16,05 ºC	3	340
Ciudad de México	33,00 ºC	17,62 ºC	12	346
Durango	36,00 ºC	17,46 ºC	97	341
Del Bajío	39,00 ºC	20,45 ºC	149	363
Oaxaca	41,00 ºC	20,56 ºC	57	363
Zacatecas	34,00 ºC	16,33 ºC	31	341
Aguascalientes	35,00 ºC	17,91 ºC	50	355
Ciudad Victoria	42,00 ºC	24,59 ºC	240	337
Retalhuleu	35,00 ºC	27,04 ºC	323	365
Querétaro	37,00 ºC	17,84 ºC	52	354
Taos	33,00 ºC	8,14 ºC	48	149
Cuernavaca	38,00 ºC	23,65 ºC	178	364
Monterrey	45,00 ºC	24,09 ºC	222	324
Saltillo	40,00 ºC	20,53 ºC	158	315
El Paso	42,00 ºC	19,46 ºC	156	264
Chihuahua	40,00 ºC	18,67 ºC	179	297
Reynosa	40,00 ºC	24,59 ºC	221	330

Taula 10. Dades atmosfèriques històriques segons ciutat de referència

Aquestes equacions van ser publicada en la revista científica *'Ecological Modelling'* [22]. La ràtio mitja anual s'obté amb l'equació:

$$Ràtio\ Mitja\ Anual = \sum_{Dia=1}^{365} \frac{Hores\ Sol\ (Dia, Latitud)}{365 \cdot 24}$$

- **Dades atmosfèriques històriques,** les variables a tenir en compte són, temperatura mitja anual, temperatura màxima, dies amb temperatures superiors als 30ºC i dies amb temperatura superior als 20ºC. Es trien aquestes variables, ja que són les que es van usar en la primera fase del projecte per tal de valorar-ne la seva viabilitat. Els valors d'aquestes variables s'assignen en funció de la seva classificació de 'Zona Climàtica', aquesta classificació assigna 19 observatoris meteorològics depenent de la localització de les EB, com es pot veure en la Taula 10, les dades es van extreure de la base de dades de la pàgina web *Weather Underground* [23].

7.3. Taula base d'anàlisi

Un cop integrades i enriquides totes les dades s'obté la taula base d'anàlisi que conté totes les dades d'interès, tots els anàlisis o operacions es realitzaran a partir d'ella. En la taula base hi ha les dades amb el major nivell de granularitat que es podrà analitzar.

Finalment la taula base d'anàlisi considerada conté 1039 files, una per a cada EB pertanyent al projecte, i 196 columnes amb diferents variables, els grups de variables es poden observar a continuació.

Grups de variables tingudes en compte

- **Info**. Aquest grup conté tota la informació de tipologia de les EB
- **Temp_ext**. Aquest grup conté els valors de temperatura exterior captats pels sensors amb una agregació mensual.
- **Temp_int**. Aquest grup conté els valors de temperatura interior captats pels sensors amb una agregació mensual.
- **is_LTE**. Conté la data d'instal·lació, l'increment de consum i la ràtio de dies amb instal·lació amb una agregació mensual.
- **LB**. Aquest grup conté els consums establerts coma línea base en tots els mesos de l'any.
- **Consum_CFE**. Conté els consums captats per CFE amb una agregació mensual.
- **Consum_WTB**. Conté els consums captats pel comptador instal·lat i enviats a WTB, amb una agregació mensual.
- **Consum_LTE**. Conté el consum mensual associat a cada EB degut a la instal·lació de nous equips.
- **Estalvi_noLTE**. Aquest grup conté el percentatge d'estalvi global de cada EB (sense tenir en compte l'increment LTE) i el valor d'energia estalviada amb una agregació mensual.

8. Anàlisi preliminar

L'anàlisi es divideix en dues parts amb diferents objectius. En la primera s'intenta millorar els resultats d'estalvi que s'estan obtenint en el projecte. En canvi en la segona part es vol extreure coneixement, trobant quines influències tenen les respectives variables en el funcionament de la solució FC.

Seguint la metodologia proposada s'extreuen un seguit de conclusions de l'anàlisi, aquestes conclusions es plasmaran immediatament després de cada apartat per tal de facilitar-ne la lectura i comprensió.

8.1. Estudi d'estalvis

ETK té coneixement de la problemàtica que suposa l'addició d'equips en els centres, però actualment no és fàcil veure l'afectació individual que suposa aquesta problemàtica, ja que la correcció dels estalvis es fa amb les dades agregades de tots els centres.

Per veure aquesta afectació s'ha recalculat, per a cada EB, el percentatge d'estalvi real, tenint en compte la correcció deguda a l'addició d'equips.

Un cop realitzat el recàlcul es compara el percentatge d'estalvi real amb el percentatge d'estalvi que no té en compte l'addició d'equips. Per fer aquesta comparació es separen els centres en dos grups, el grup amb les EB que tenen informació sobre l'addició d'equips i el grup que no té cap tipus d'informació. Com es pot veure en la Figura 16 el primer grup passa d'un 6,96% d'estalvi a un 22%, mentre que en els centres del segon grup l'estalvi passa d'un 24,6% a un 24,9% (aquesta diferència del 0,3% s'associa a l'ús de dades agregades durant el nou càlcul d'estalvis, ja que teòricament no hauria de variar). D'aquesta comprovació es confirma la gran influència que té l'addició d'equips en el percentatges d'estalvi obtingut.

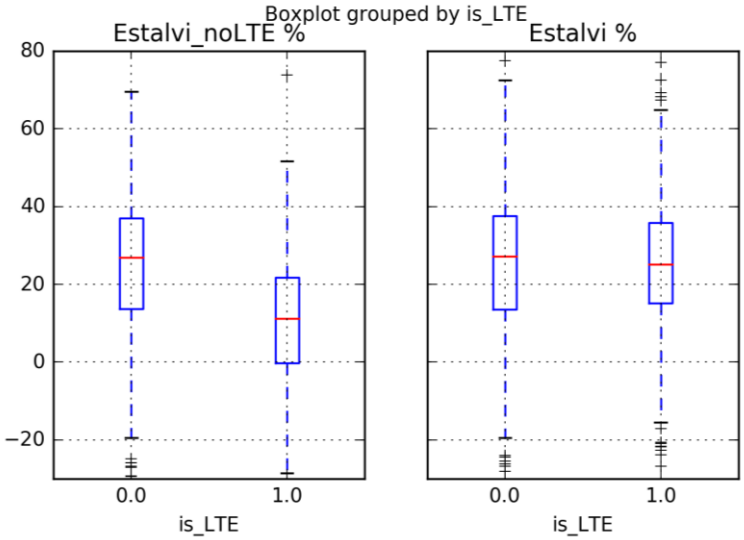


Figura 16. BoxPlot del percentatge d'Estalvi_noLTE i d'Estalvi real agurpats per existència d'informació sobre addició de equips LTE

Actualment no hi ha intercanvi d'informació, sobre l'addició d'equips, entre TEF i ETK, es sospita que hi pot haver un nombre important de centres sense informació que han patit una addició d'equips. Veient la gran afectació que suposa l'addició d'equips en l'obtenció de l'estalvi objectiu i per tant en el retorn econòmic (com s'ha comentat en l'apartat 5.1, les empreses del tipus ESCO facturen en funció de l'estalvi energètic generat). S'haurien d'incrementar els esforços per obtenir aquesta informació, però mentre aquesta no s'obtingui aquesta, ETK hauria de centrar part dels seus esforços en detectar remotament aquesta addició.

8.1.1. Script de detecció d'increments de consum

Per tal de realitzar la tasca de detecció d'una manera més automatitzada s'ha creat un script que a partir de les dades emmagatzemades en la taula base d'anàlisi avalua els gradients de consum diari entre mesos. Com a resultat de l'execució de l'*script* s'obté un llistat de les EB que són sospitoses d'haver patit addició d'equips. A continuació es comenten tres casos genèrics detectats com a conseqüència de l'execució de l'script, per detectar increments superiors als 16 kWh/dia. Aquest valor s'obté a partir de l'increment promig detectat en els centres amb informació.

Addició d'equips

En la Figura 17 es pot observar un exemple ideal d'addició de equips en la EB *El Ruso*. Al tractar-se d'una EB amb un consum baix, l'increment de consum percentual és molt elevat i fàcil de veure. El primer increment del consum que es pot veure correspon a un període de prova de l'AA, ja que durant aquell període el ventilador deixa de funcionar.

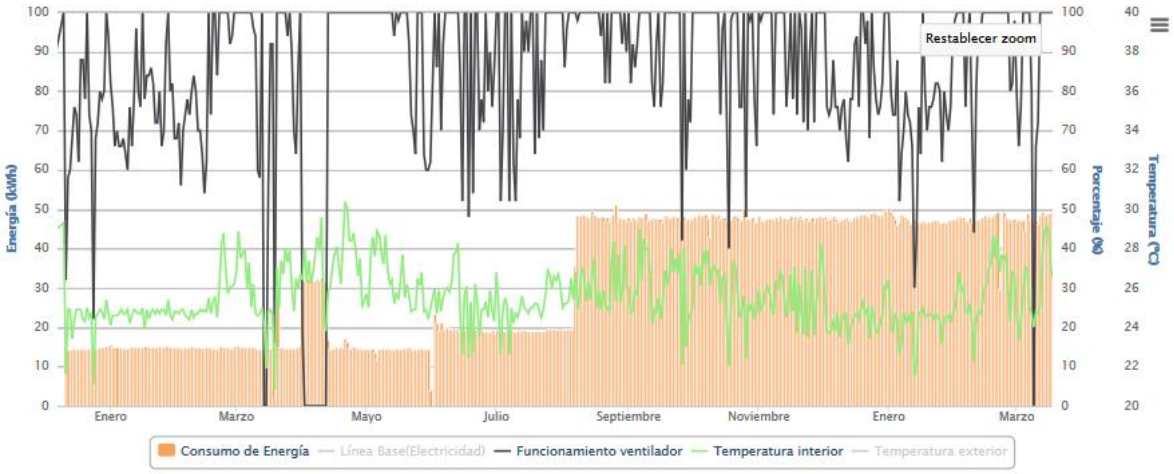


Figura 17. Evolució de la EB EL Ruso (15-03062)

En canvi després es poden observar dos increments molt marcats. Aquests increments són deguts a l'addició d'equips, el segon s'associa a l'addició del LTE, en canvi el primer s'associa hipotèticament a millores en els equips UMTS. Aquesta hipòtesi s'extreu a partir de la informació aportada per part del personal mantenidor de Guatemala, que senyala la recurrència entre la millora d'equips UMTS i la instal·lació LTE amb entre 1 i 3 mesos de diferència.

Contactors modificats

En la Figura 18 es pot observar un bon exemple de modificació de contactors, és a dir, anul·lar l'efecte dels contactors fent una conexió en pont. Al anul·lar l'efecte dels contactors l'equip AA disposa d'electricitat en tot moment, i si aquest té una temperatura de consigna baixa, com en el cas de la Figura 18 provoca que la lógica del sistema FC no entri mai en funcionament, deixant els ventiladors permanentment parats. Aquesta casuística s'ha donat en multitud de centres i és degut al desconeixement del sistema FC per part del personal mantenedor dels AA, i a la manca de comunicació per part de TEF. Els pics de funcionament dels ventiladors, que es poden observar mentre els contactors estan modificats, s'associen a entrades de personal de manteniment a la EB amb el consegüent increment de la temperatura interior i encesa temporal dels ventiladors.

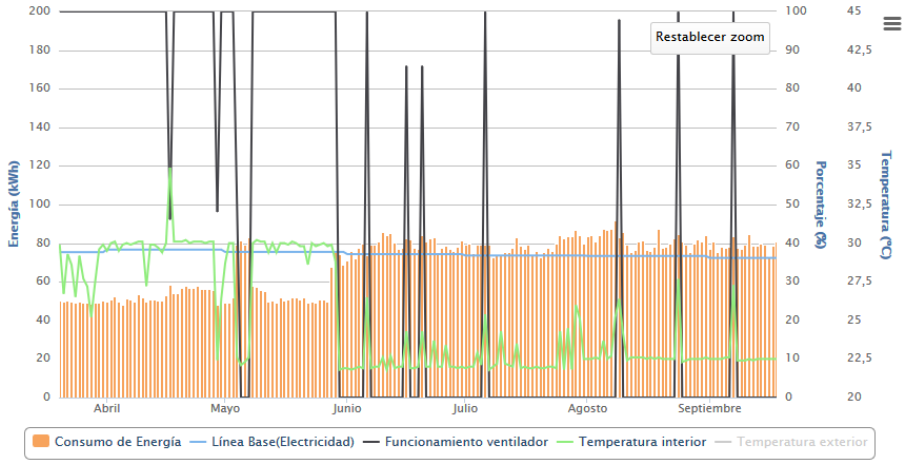


Figura 18. Evolució de la EB Xonacatepec (21-03095)

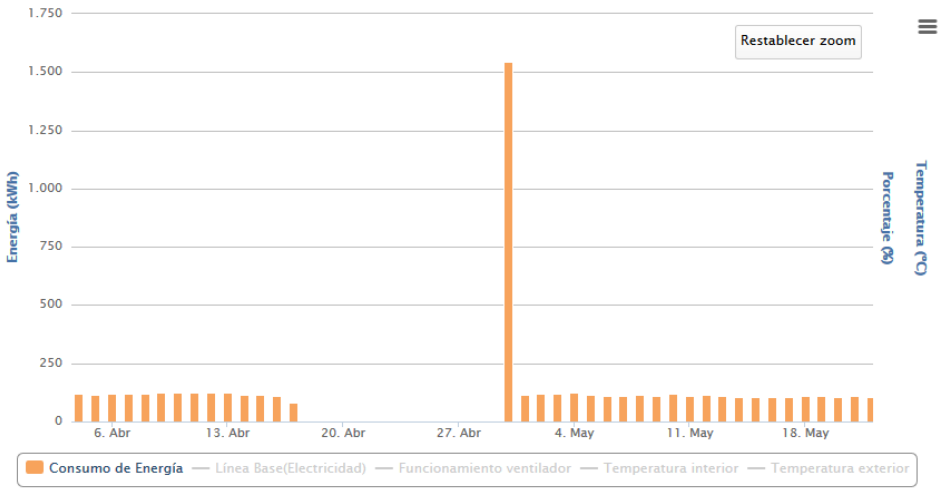


Figura 19. Exemple de fals positiu EB Doroteo Arango (09-03318)

Fals positiu

L'explicació de la majoria de falsos positius és la següent, de tant en tant el mòdem és incapaç de enviar les dades recollides pel PLC, per exemple degut a la falta de cobertura, això provoca que en el moment en que es restableix la comunicació es mostri tota la energia consumida els dies anteriors, com es pot veure en la Figura 19. Aquesta casuística es dóna perquè la variable *Energia Consumida*, és una variable continua, és a dir, envia l'energia consumida acumulada de la EB. Si la pèrdua de comunicació no és molt prolongada la plataforma WTB no és capaç de detectar aquests increments de consum.

8.1.2. Conclusions

Amb l'execució d'aquest *script* es van detectar 39 centres sospitosos d'haver patit una important pujada de consums. D'aquests 17 centres havien patit una presumpta addició d'equips, 12 se'ls havia modificat els contactors, mentre que els 10 restants corresponen a falsos positius.

Veient el nombre de centres en que s'ha detectat una modificació de contactors es pot concloure que és una problemàtica real que caldria resoldre, ja que anul·la els estalvis obtinguts gràcies al FC. Per tant s'hauria d'avaluar la viabilitat i necessitat de destinar recursos per solucionar-ho.

Amb un encert en la detecció d'increment de consum del 43,6% i un 30,8% en la detecció de contactors manipulats, s'aconsegueix obtenir informació útil d'un 74,4% de les EB del llistat de centres sospitosos. Els resultats obtinguts es valoren molt positivament tenint en compte la simplicitat del *script* de detecció i l'alt nivell d'agregació de les dades usades.

Veient el percentatge d'èxit obtingut s'hauria d'avaluar la implementació d'un *script* de detecció més elaborat i que treballés amb una agregació menor, d'aquesta manera es podrien evitar falsos positius i detectar variacions menors. Es creu viable per dos motius, en primer lloc es sospita que es podria detectar una quantitat important de centres, i en segon lloc perquè aquesta problemàtica pot ser recurrent durant els 6 anys de duració del projecte.

8.2. Correlacions entre variables

Per tal de localitzar certs patrons, relacions fortes o inusuals es decideix correlacionar totes les variables de la taula base d'anàlisi. Al aplicar aquesta correlació s'obté una matriu de correlació de 196x196, la magnitud de la matriu fa que sigui molt difícil extreure'n algun tipus d'informació, per facilitar aquest procés es crea un mapa de calor, aquest mapa es pot veure en la Figura 20. Un mapa de calor és una representació gràfica dels valors individuals continguts en una matriu, en cada cel·la s'hi assigna un color, contingut en un espectre de colors. El color es tria en funció del seu valor, el percentil al qual pertany, o la propietat desitjada. En aquest cas el blau correspon a valors superiors a 0,5 , el vermell a valors inferiors -0,5 , mentre que els valors compresos entre 0,5 i -0,5 són una interpolació entre aquests i el color groc que pertany al valor 0.

Alhora d'avaluar aquest mapa de calor és interessant fixar-se en les zones calentes i fredes que no es troben en la diagonal, ja que en la diagonal les correlacions són molt elevades degut a que s'estan correlacionant el mateix tipus de variables, per exemple, consum amb consum.

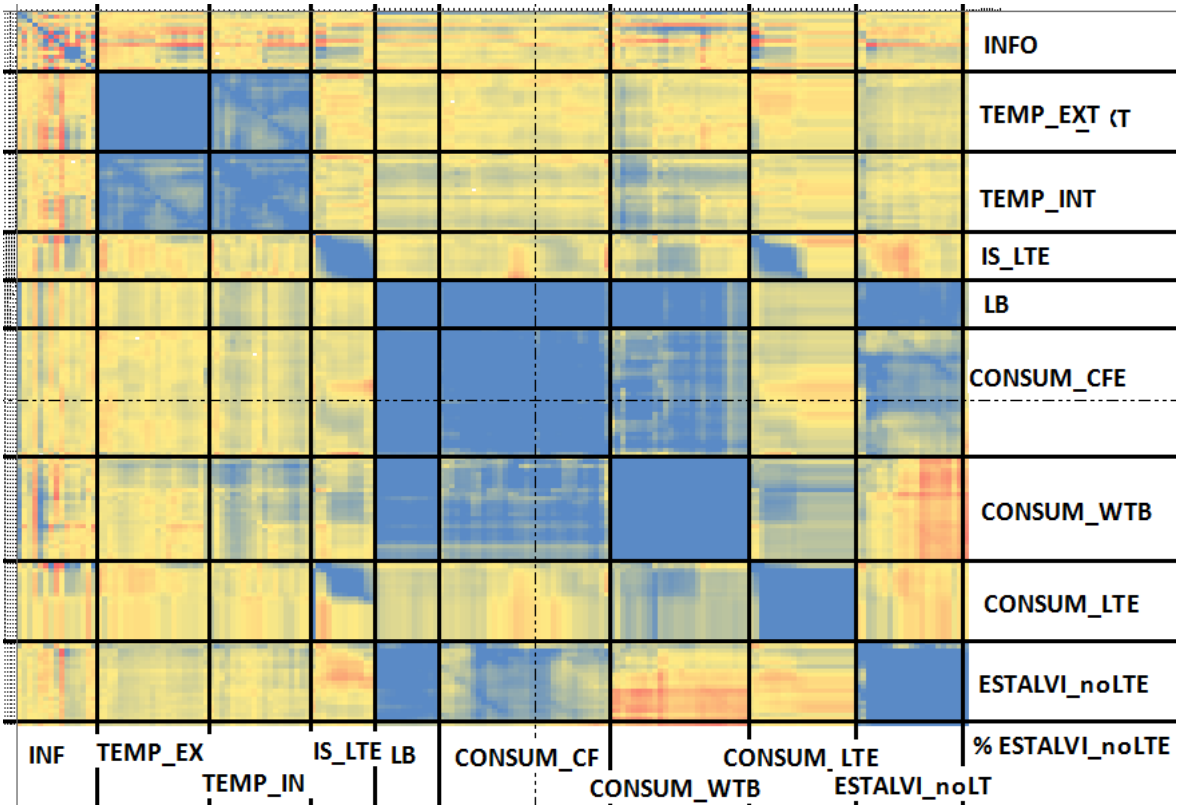


Figura 20. Mapa de calor de la matriu de correlacions

Per tal de simplificar l'anàlisi es disminueix la granularitat de les dades, creant dues noves taules. En les dues s'hi tenen en compte els últims nou mesos perquè són els que contenen menys *missing*. La raó per la qual hi ha més valors vàlids és simple, en aquella data s'havia executat la instal·lació de la totalitat de les EB. En la primera taula s'hi té en compte una granularitat temporal trimestral, mentre que en la segona s'estableix una granularitat de nou mesos per aconseguir obtenir una única columna per tipus de variable. En aquestes noves taules també si correlacionen les variables, es troben en la Figura 21 i la Figura 22.

Cal tenir en compte que aquestes correlacions són una primera aproximació a les relacions entre les variables i no tenen un rigor estadístic molt elevat, ja que les dades contenen un biaix important. Tot i així poden aportar relacions interessants que caldria estudiar amb més profunditat.

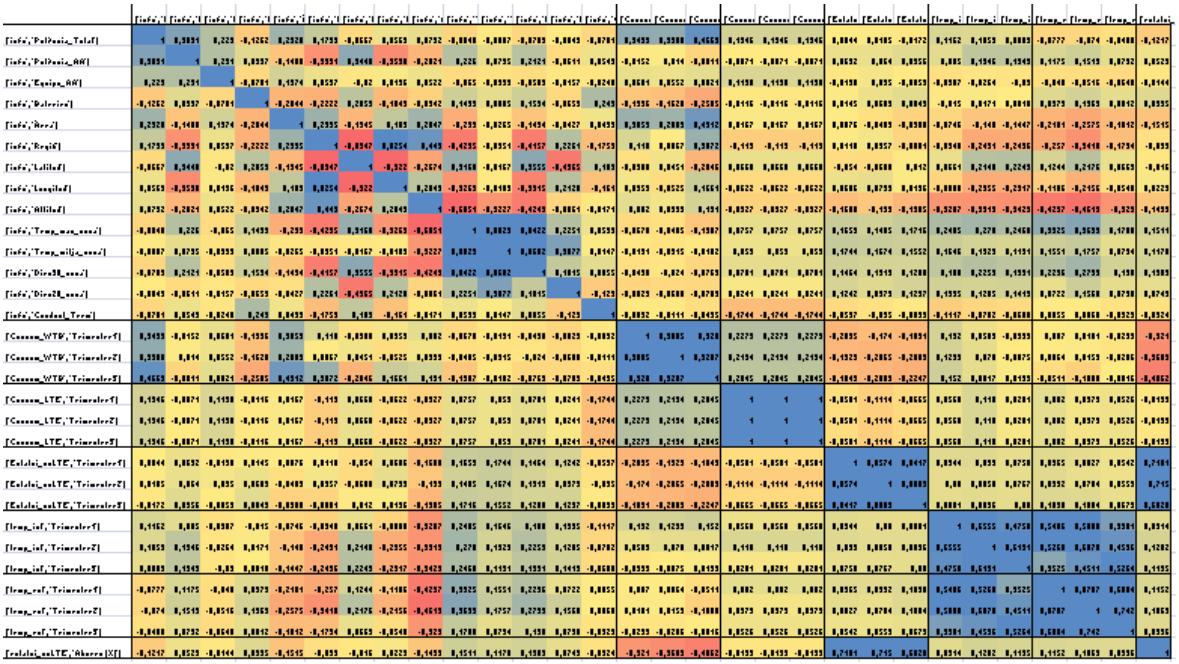


Figura 21. Mapa de calor de la matriu de correlacions amb agregació trimestral

	('info', 'Potència_Total')	('info', 'Potència_AA')	('info', 'Equips_AA')	('info', 'Bateries')	('info', 'Àrea')	('info', 'Regió')	('info', 'Latitud')	('info', 'Longitud')	('info', 'Altitud')	('info', 'Temp_max_zona')	('info', 'Temp_mitja_zona')	('info', 'Dies30_zona')	('info', 'Dies20_zona')	('info', 'Conduct_Term')	Consum_WTB	Consum_LTE	Estalvi_noLTE	temp_int	temp_ext	('estalvi_noLTE', 'Ahorro (%)')
('info', 'Potència_Total')	1	0,3831	0,229	-0,126	0,2928	0,1733	-0,067	0,0563	0,0732	-0,085	-0,009	-0,078	-0,005	-0,07	0,3371	0,1346	0,0053	0,0968	-0,077	-0,122
('info', 'Potència_AA')	0,3831	1	0,231	0,0937	-0,149	-0,339	0,3448	-0,36	-0,202	0,226	0,0795	0,2121	-0,061	0,0543	-0,001	-0,007	0,0622	0,1751	0,1165	0,0529
('info', 'Equips_AA')	0,229	0,231	1	-0,07	0,1374	0,0597	-0,02	0,0196	0,0522	-0,065	-0,034	-0,051	-0,016	-0,025	0,0609	0,1138	0,0029	-0,037	-0,059	-0,014
('info', 'Bateries')	-0,126	0,0937	-0,07	1	-0,204	-0,222	0,2059	-0,185	-0,094	0,1493	0,0805	0,1534	-0,066	0,249	-0,158	-0,012	0,0504	0,0042	0,1077	0,0995
('info', 'Àrea')	0,2928	-0,149	0,1374	-0,204	1	0,2995	-0,195	0,183	0,2847	-0,233	-0,026	-0,143	-0,043	0,0439	0,3012	0,0167	-0,033	-0,138	-0,219	-0,152
('info', 'Regió')	0,1733	-0,339	0,0597	-0,222	0,2995	1	-0,835	0,8254	0,449	-0,423	-0,035	-0,416	0,2261	-0,175	0,049	-0,113	0,0258	-0,256	-0,284	-0,039
('info', 'Latitud')	-0,067	0,3448	-0,02	0,2059	-0,195	-0,835	1	-0,922	-0,267	0,3168	-0,017	0,3555	-0,436	0,183	0,0026	0,0668	-0,063	0,2167	0,168	-0,016
('info', 'Longitud')	0,0563	-0,36	0,0196	-0,185	0,183	0,8254	-0,922	1	0,2049	-0,327	-0,01	-0,392	0,2128	-0,161	-0,003	-0,062	0,0767	-0,232	-0,163	0,0223
('info', 'Altitud')	0,0732	-0,202	0,0522	-0,094	0,2847	0,449	-0,267	0,2049	1	-0,605	-0,323	-0,424	-0,086	-0,017	0,0564	-0,093	-0,154	-0,395	-0,437	-0,144
('info', 'Temp_max_zona')	-0,085	0,226	-0,065	0,1493	-0,233	-0,423	0,3168	-0,327	-0,605	1	0,8023	0,8422	0,2251	0,0539	-0,063	0,0757	0,1513	0,31	0,3187	0,1511
('info', 'Temp_mitja_zona')	-0,009	0,0795	-0,034	0,0805	-0,026	-0,035	-0,017	-0,01	-0,323	0,8023	1	0,8682	0,3877	0,0147	-0,029	0,053	0,1674	0,1646	0,1477	0,1178
('info', 'Dies30_zona')	-0,078	0,2121	-0,051	0,1534	-0,143	-0,416	0,3555	-0,392	-0,424	0,8422	0,8682	1	0,1815	0,0855	-0,034	0,0781	0,1328	0,2499	0,2349	0,1383
('info', 'Dies20_zona')	-0,005	-0,061	-0,016	-0,066	-0,043	0,2261	-0,436	0,2128	-0,086	0,2251	0,3877	0,1815	1	-0,123	-0,09	0,0241	0,1077	0,1609	0,113	0,0749
('info', 'Conduct_Term')	-0,07	0,0543	-0,025	0,249	0,0439	-0,175	0,183	-0,161	-0,017	0,0539	0,0147	0,0855	-0,123	1	-0,03	-0,174	-0,04	-0,066	0,0037	-0,032
Consum_WTB	0,3371	-0,001	0,0609	-0,158	0,3012	0,049	0,0026	-0,003	0,0564	-0,063	-0,029	-0,034	-0,09	-0,03	1	0,224	-0,186	0,0486	0,0047	-0,333
Consum_LTE	0,1346	-0,007	0,1138	-0,012	0,0167	-0,113	0,0668	-0,062	-0,093	0,0757	0,053	0,0781	0,0241	-0,174	0,224	1	-0,086	0,093	0,0673	-0,019
Estalvi_noLTE	0,0053	0,0622	0,0029	0,0504	-0,033	0,0258	-0,063	0,0767	-0,154	0,1513	0,1674	0,1328	0,1077	-0,04	-0,186	-0,086	1	0,095	0,072	0,7433
temp_int	0,0968	0,1751	-0,037	0,0042	-0,138	-0,256	0,2167	-0,232	-0,395	0,31	0,1646	0,2499	0,1609	-0,066	0,0486	0,093	0,095	1	0,6502	0,1335
temp_ext	-0,077	0,1165	-0,059	0,1077	-0,219	-0,284	0,168	-0,163	-0,437	0,3187	0,1477	0,2349	0,113	0,0037	0,0047	0,0673	0,072	0,6502	1	0,1097
('estalvi_noLTE', 'Ahorro (%)')	-0,122	0,0529	-0,014	0,0995	-0,152	-0,039	-0,016	0,0223	-0,144	0,1511	0,1178	0,1383	0,0749	-0,032	-0,333	-0,019	0,7433	0,1335	0,1097	1

Figura 22. Mapa de calor de la matriu de correlacions amb agregació de 9 mesos

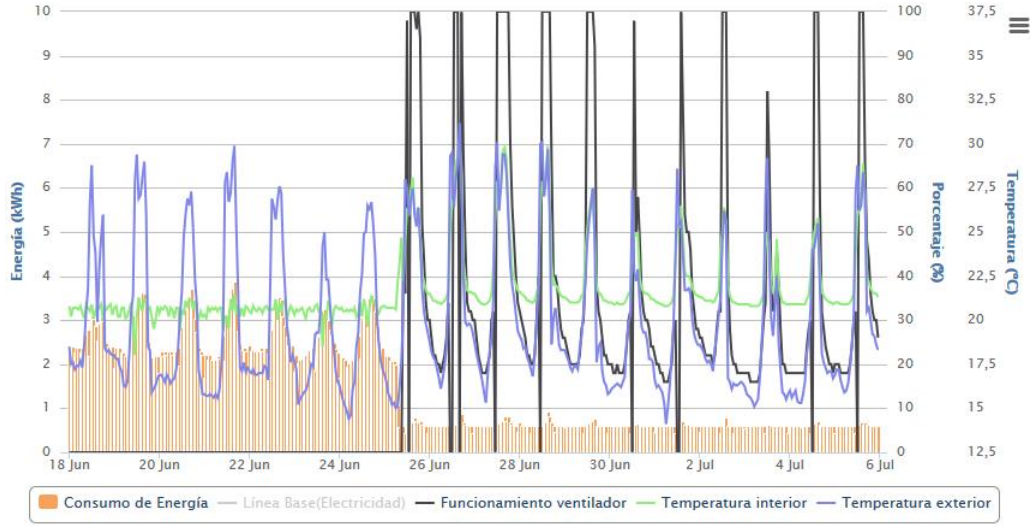


Figura 23. EB amb AA sobredimensionat i gran potencial d'estalvi (Durango 22-03215)

8.2.1. Relacions associades al percentatge d'estalvi

8.2.1.1. Potència

Es pot observar que la variable *Potència* està correlacionada negativament amb el percentatge d'estalvi, la seva correlació és de -0,122. A continuació es formula la hipòtesi sobre la correlació obtinguda.

La gran majoria d'equips AA tenen una potència elèctrica d'entre 1,5 i 2 kW i en la majoria d'EB només n'hi ha una unitat instal·lada. D'aquest fet se'n pot extreure que en les EB amb un consum baix, els equips AA estan sobredimensionats. Per exemple si es té en compte que l'EER[4] és 3 un equip pot extreure entre 4,5 i 6 kW de potència tèrmica, mentre que hi ha un bon nombre de EB amb una potència demandada menor a 4kW. Aquestes EB són les que tenen un potencial d'estalvi major, la Figura 23 n'és un exemple ideal.

En canvi en les EB amb un consum elevat no és usual que els equips AA estiguin sobredimensionats, això disminueix el potencial d'estalvi, sobretot si es té amb compte que s'està avaluant l'estalvi percentual. A això cal afegir-hi que en alguns casos d'equips amb consum molt elevat disposen d'un extractor propi, disminuint el percentatge del possible estalvi.

[4] Energy Efficiency Ratio, és la ràtio que s'usa per valorar la eficiència d'un equip que treballa en un cicle de refredat, es calcula de la següent manera: $EER = (Potència\ Tèrmica\ Extreta) / (Potència\ Elèctrica\ Consumida)$

Correlació	Temp_max_zona	Temp_mitja_zona	Temp_int_WTB	Temp_ext_WTB
Estalvi (%)	0,151	0,118	0,134	0,110
Altitud	-0,605	-0,323	-0,395	-0,437

Taula 11. Taula de correlacions entre l'Altitud, Percentatge d'estalvi i les Temperatures

8.2.1.2. Altitud i temperatures

Es pot observar que la variable *Altitud* està correlacionada negativament amb el percentatge d'estalvi, la seva correlació és de -0,144, mentre que les correlacions amb les variables de temperatura són positives, el seus valors es poden observar en la Taula 11. Cal destacar que les variables enriquides referents a la temperatura històrica tenen una influència important.

Les correlacions entre l'*Altitud* i les *Temperatures* són molt fortes, aquesta correlació és totalment lògica, i es poden veure en la Taula 11.

Veient aquestes relacions es pot extreure que la temperatura és una variable que afecta de manera important en els estalvis obtinguts, i a la vegada, com era d'esperar, l'altitud té una afectació molt elevada en la temperatura exterior.

Explicació de les relacions entre la temperatura exterior i l'estalvi potencial

La relació entre la temperatura exterior i l'estalvi obtingut no és, en cap cas, lineal. Per una banda a major temperatura hi ha un major potencial d'estalvi degut a que abans de la instal·lació del FC l'equip AA funcionava durant més temps amb una eficiència menor, això implica que un major percentatge del consum de la LB correspon al sistema de refrigeració. Però per l'altra banda si la temperatura és molt elevada durant gran part del dia, el sistema FC romandrà inactiu durant aquell temps, és a dir, sense reduir el consum anterior. a la instal·lació del FC.

8.2.1.3. Àrea

Es pot observar que la variable *Àrea* està correlacionada negativament amb el percentatge d'estalvi, la seva correlació és de -0,152.

La correlació entre l'*Àrea* i el *Consum WTB* monitorat és de 0,303 , aquest valor sembla totalment lògic, ja que usualment les EB més grans contenen un major nombre d'aparells i conseqüentment un major consum. Aquesta relació es pot comprovar en el BoxPlot de la Figura 24, en el qual s'ha discretitzat l'àrea en tres grups i s'hi ha comparat les potències associades (l'eix y s'ha limitat a 9 kW per aconseguir una visió més clara, provocant que 12 EB quedin excloses del gràfic).

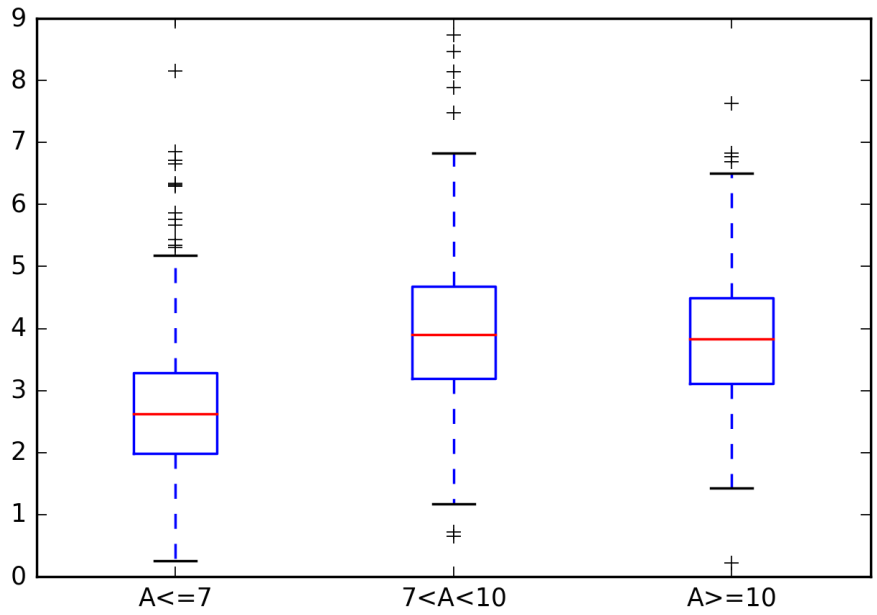


Figura 24. BoxPlot de la potència monitorada (kW) en funció de l'àrea (m^2)

8.2.1.4. Conculsions

En cas de expandir el projecte FC a altres regions i/o països, seria interessant prosseguir amb major rigor i profunditat l'anàlisi de relacions entre variables de tipologia i d'estalvi. D'aquesta manera es podria triar els centres amb major potencial d'estalvi, disminuint el nivell d'incertesa del projecte. Es podria tenir una major seguretat de cara a l'obtenció de l'estalvi objectiu, i per tant disminuir el període de retorn del projecte.

8.2.2. Relacions peculiars

8.2.2.1. Capacitat de les bateries

S'han detectat dues relacions peculiars amb la variable *Capacitat de les bateries*. La primera d'elles és la correlació negativa de -0,158 amb el *Consum* i la segona és la correlació negativa de -0,204 amb l'*Àrea*. Aquesta relació s'ha classificat com a peculiar perquè les EB amb major consum haurien de disposar d'una major capacitat de bateries.

S'hauria d'estudiar més a fons l'origen d'aquesta correlació, per descartar que aquesta provingui d'altres relacions. Per exemple del nivell d'afectació de la caiguda de la EB, és a dir, que les EB més crítiques, de cara a mantenir el servei de telecomunicacions, tinguin una major capacitat de bateries; o que les EB amb major consum i àrea disposin d'un generador d'electricitat auxiliar. En cas de descartar aquest tipus de factors la relació indicaria que la disposició de bateries en les EB no és la òptima i/o adequada.

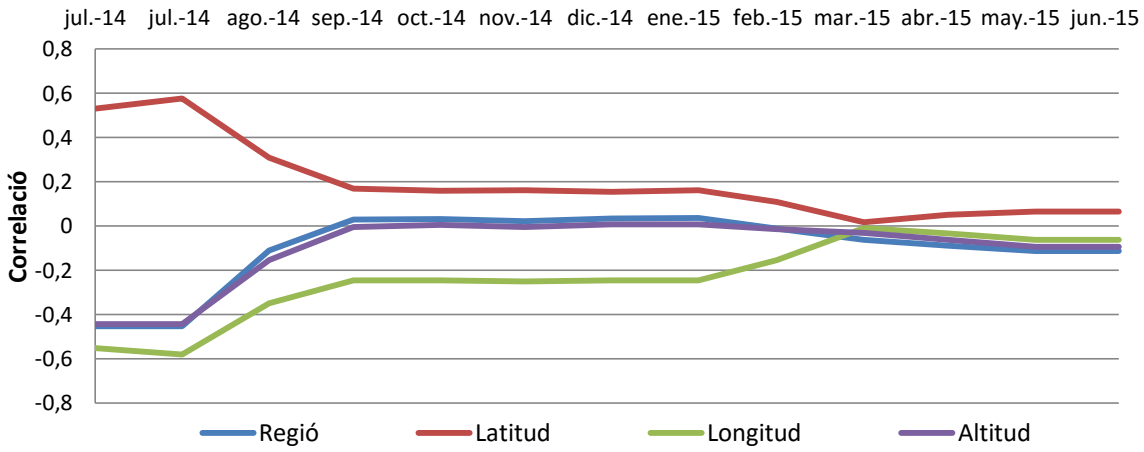


Figura 25. Evolució temporal de la correlació entre el consum degut a LTE i les variables: Regió, Latitud, Longitud i Altitud

8.2.2.2. Consum LTE

En els primers mesos s'han detectat unes correlacions molt fortes entre la variable *Consum LTE* i les variables geogràfiques. Aquest fet és degut al biaix que hi ha en les dades, aquest biaix es justifica amb dos factors. El primer és el baix nombre d'E B en que s'havia instal·lat equips LTE, en comparació a els últims mesos. El segon és la planificació d'instal·lació d'equips ideada per TEF que amb tota probabilitat té un fort component geogràfic de cara al desplegament de la tecnologia. En la Figura 25 es pot observar com a mesura que s'instal·la LTE en un major nombre de EB les correlacions passen a ser molt properes a zero, d'aquest fet s'extreu que el baixi introduït disminueix considerablement i que no hi ha cap relació entre el consum degut als nous equips respecte les variables geogràfiques.

9. Avaluació, recomanacions i desplegament

9.1. Avaluació

Com a avaluació general, es conclou que s'han assolit els objectius proposats en l'inici del projecte (2.3), ja que s'ha desenvolupat i posat a prova una metodologia que ha mostrat ser capaç de generar resultats molt positius. Fins i tot amb una inversió de recursos personals i temporals molt baixa (des del punt de vista d'una organització) i únicament fent ús de d'eines i tècniques analítiques bàsiques. A la vegada també és una metodologia flexible, ja que es podria estendre fàcilment aplicant anàlisis més elaborats sobre la taula base d'anàlisi ja obtinguda, això afavoreix la possibilitat de realitzar anàlisis més profunds sense la necessitat de fer una gran inversió.

Després d'aplicar la metodologia proposada en el projecte Free-Cooling Mèxic, com a prova de concepte, es conclou que la implementació d'un nivell d'anàlisi més elevat en el conjunt de l'empresa, no només és viable, sinó que també és desitjable des del punt de vista empresarial. Per avaluar-ne la viabilitat s'ha tingut en compte el basant tècnic, l'econòmic i el potencial de resultats positius.

Els resultats de l'anàlisi preliminar i les seves conclusions demostren el potencial d'obtenir coneixements amb capacitat d'influir en els bons resultats dels projectes, i per tant disminuir-ne el retorn. Alguns d'aquests resultats podrien ser quasi instantanis, com la detecció de mal comportament en les EB. També facilitaria la realització d'accions de l'operativa diària de l'empresa, ja que en alguns casos les dades requereixen d'una neteja i preparació prèvia, que ja no seria necessària. Per altra banda, i a més llarg termini, també es podrien estudiar els paràmetres de funcionament que optimitzen la solució, o la tipologia de EB amb major potencial d'estalvi, de cara a expandir aquest projecte a altres regions.

Per altra banda, actualment no és necessari fer ús de les eines desenvolupades per les grans empreses del sector. Aquestes eines propietàries tenen un cost bastant elevat, que moltes organitzacions no poden assumir, i a la vegada tenen una corba d'aprenentatge més llarga. Tal i com ha quedat demostrat és totalment factible l'ús d'eines *free and open source*. Aquestes eines obren les portes a moltes organitzacions, per començar projectes de DM, perquè la inversió a realitzar és menor i per tant és fàcil disminuir el retorn d'aquesta. També aporten una major grau de flexibilitat i modularitat a l'hora de tractar les dades, bàsicament per dues raons, usualment permeten ampliar les seves funcions fent ús d'altres llibreries i

també permeten exportar les dades amb formats estàndard que són reconeguts per la resta d'eines usades per al Data Mining.

9.2. Recomanacions

- Automatitzar la integració de dades que es creen, per tenir la taula d'anàlisi actualitzada i disminuir el temps dedicat a la integració.

- Integrar manualment dades emmagatzemades en antics excels de funcionament operatiu. Ja que, actualment no se'n fa ús i no es possible fer-ne un anàlisi. De la integració i el posterior anàlisis sobre aquestes dades, actualment disperses, es podrà extreure un cert valor.

- Per expandir aquest model d'anàlisi en els altres projectes de l'empresa seria necessari dedicar personal a temps complert o parcial, per tal de controlar la integració de dades, realitzar anàlisis i subministrar les dades als departaments que ho necessitessin.

- S'ha de plantejar la contractació d'un expert en tècniques de modelatge. Aquest fet ve motivat per la possibilitat de realitzar anàlisis més profunds per poder extreure més informació subjacent a les dades.

9.3. Desplegament

Com a resultat de presentar a l'empresa la problemàtica del volum d'informació de tipologia, que s'emmagatzemava en excels operatius durant el període d'execució i que actualment no se'n feia cap ús, s'ha decidit crear *tags* a la plataforma Wattabit. Els *tags* són unes etiquetes que descriuen la EB, s'ha començat assignant etiquetes en funció dels blocs, regions, potències, mode de funcionament i informació sobre LTE. L'arbre jeràrquic d'aquests *tags* es pot veure en la Figura 26. La realització d'aquest integració de dades, provinents d'excels operatius, a la plataforma Wattabit és una mostra de la presa de consciència, per part de l'empresa, de la importància d'emmagatzemar les dades allà on puguin ser d'utilitat.

S'ha planificat i desenvolupat la creació d'una eina d'extracció massiva de les dades contingudes en la plataforma Wattabit. Aquesta acció s'ha realitzat com a conseqüència de presentar, en el Resum de situació (4.2.3), la gran afectació negativa que suposava no disposar d'aquesta eina. El fet de donar solució a una de les problemàtiques més grans que s'havien detectat, es valora molt positivament, ja que demostra l'interès de l'empresa en l'anàlisi de dades. En la Figura 26 es pot veure la interfície gràfica que permet realitzar extraccions, a Excel o CSV, amb diferents granularitats.

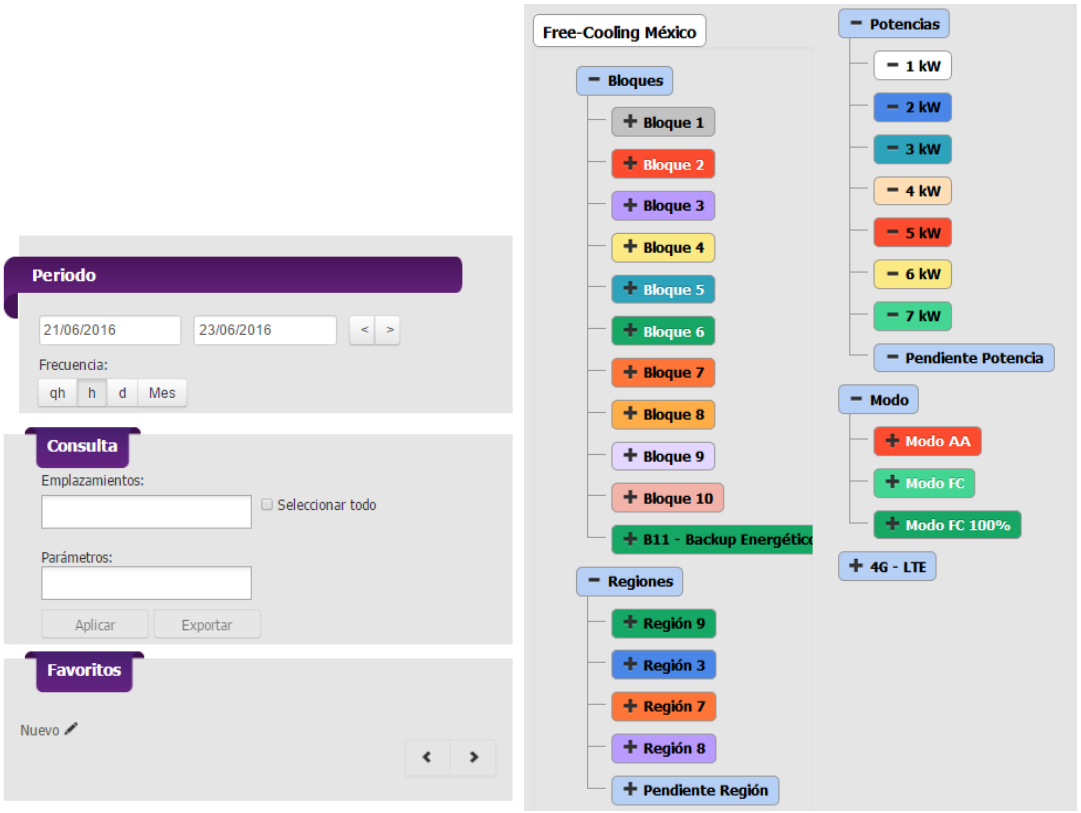


Figura 26. Arbre jeràrquic dels tags (dreta) i eina d'extracció de dades (esquerra)

Per últim s'ha planificat la modificació de les alarmes de manera que se'n pugui emmagatzemar un històric per EB i tipologia, i a la vegada guardar els valors llindar usats en cada període. D'aquesta manera s'evitarà perdre informació referent a les alarmes i a la vegada incrementar-ne la seva fiabilitat i utilitat.

10. Eines usades

En el sector del DM existeixen multitud d'eines i moltes vegades per un mateix projecte se n'usen diverses, la gran majoria d'eines comercials estan molt focalitzades per diferents tipus de negocis. El mercat del software propietari per a la preparació de dades la integració l'anàlisi,etcètera es podria considerar un mercat competitiu, on les 10 empreses millor posicionades *només* controlen el 50% del volum de ventes del sector.

Per a la realització d'aquest projecte, i tenint en compte que no es comptava amb el suport econòmic de l'empresa, es va elegir fer ús d'eines gratuïtes i preferiblement *open source*.

Es va realitzar un sondeig en webs especialitzades sobre les eines gratuïtes més usades. També es va tenir amb compte que aquestes tinguessin una gran flexibilitat a l'hora d'importar documents; fossin capaces de netejar, preparar i integrar les dades d'una manera relativament senzilla.

Opcions avaluades

- **R** [24]. És un entorn i llenguatge de programació enfocat a l'anàlisi estadístic i el tractament de dades. Un dels seus gran avantatges és la multitud de funcions estadístiques, analítiques i de tractament que té implementades, també cal destacar es seves funcions gràfiques. Es tracta del llenguatge més usat per la comunitat estadística i també és molt popular en la mineria de dades.

  Es va desestimar l'opció d'usar aquesta eina degut a la major corba d'aprenentatge envers altres de les opcions, ja que no només calia familiaritzar-se amb l'entorn, sinó que també calia fer-ho amb el llenguatge.

- **Pandas** [25]. En aquest cas es tracta d'una llibreria, escrita en Python, per a la manipulació i preparació de dades. Aquesta llibreria ofereix unes estructures de dades i operacions per manipular-les molt eficients. Pandas no implementa una quantitat tant elevada de model estadístics com fa R, però això es pot solucionar fent ús d'altres llibreries de Python, que estan preparades per treballar amb les estructures de Pandas.

  Aquest és el programa que s'ha usat per a la realització de totes les tasques tractament, integració i anàlisi de dades; i per a la majoria de tasques de neteja i preparació. Ha sigut la opció elegida pel fet d'estar habituat al llenguatge de programació Python, reduint així la corba d'aprenentatge al funcionament de la llibreria en qüestió.

- **Talend Open Studio** [26]. Talend és una empresa que es dedica a la creació de software propietari per la integració i tractament de dades, tot i això la empresa també desenvolupa un projecte *Open Source*, en aquest projecte s'hi poden trobar programes de integració, qualitat, neteja, gestió, etcètera.

  S'ha desestimat l'ús del Studio de Talend per a les tasques de tractament del projecte, bàsicament per dues raons, perquè està més focalitzat en treballar amb diverses bases de dades amb un flux *regular* d'informació i també per la poca documentació trobada. Tot i això s'ha fet ús del programa *Talend Data Preparation* per la neteja de les dades referents a la tipologia i informació de les diferents EB, es pot veure en la Figura 27. Aquest programa aporta una informació important a l'hora de fer la neteja per exemple: mostra diagrames de freqüència de valors, detecta tipus de dades diferents en una columna, detecta valors anòmals en funció del nom de la columna (per exemple si en la columna estats apareix un estat no Mexicà és capaç de detectar-lo), permet visualitzar els diversos patrons de caràcters en cada columna, etcètera. Un cop l'analista detecta aquestes anomalies el programa et deixa modificar-les, eliminar-les o realitzar-hi la acció que escaigui. També cal destacar que crea un *full de ruta* (es pot veure en la part esquerra de la Figura 27) de les accions de neteja realitzades per després aplicar-les en altres documents.

- **Anaconda** [27]. És una distribució *freemium open source* de *Python* i *R* que inclou un gran nombre de les llibreries més usades de *Python*. Aquesta distribució està optimitzada per processar grans quantitats de dades i realitzar-hi anàlisis. Anaconda és la plataforma líder en l'*open data science*. Tots els gràfics referents a les dades s'han realitzat a partir de la llibreria de Python **Matplotlib** [28] (continguda en el paquet de llibreries d'Anaconda) i extensions de la mateixa, com per exemple en el cas dels gràfics realitzats sobre plànols.

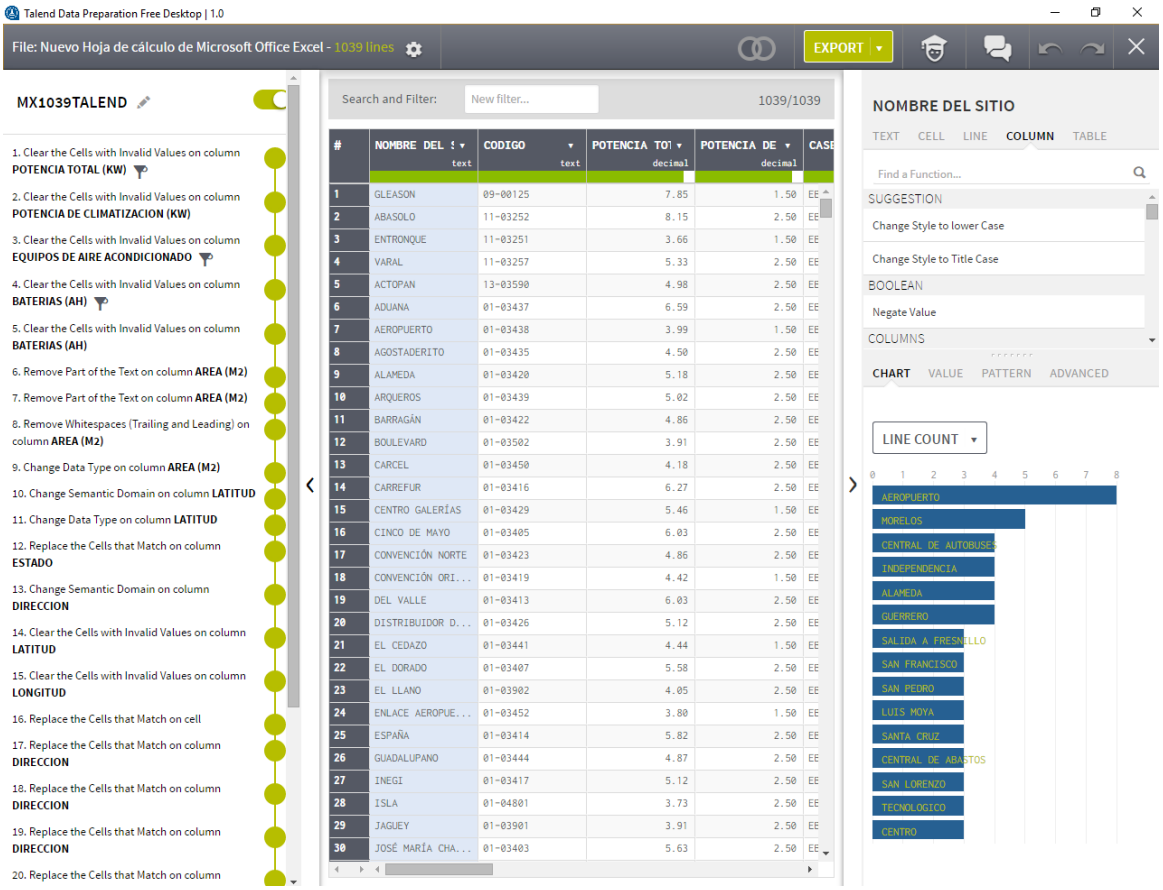


Figura 27. Imatge del Software de neteja i preparació Talend

11. Planificació temporal i costos

11.1. Planificació temporal

L'objectiu de la planificació del projecte es obtenir una distribució en el temps de les activitats que s'han de realitzar, per poder optimitzar el temps dedicat, i conseqüentment, el cost del projecte.

El projecte s'ha dividit en diverses fases, que en un principi són més difuses, i a mesura que s'avança en el projecte es poden delimitar amb tota perfecció. En aquest projecte les fases han quedat perfectament definides amb la fi de la segona fase, que marca la metodologia a seguir per fer la prova de concepte.

El projecte ha tingut una durada de 19 setmanes i s'estima que la dedicació temporal en hores ha sigut de 450. En la Taula 12 es pot veure el cronograma o diagrama de Grantt de les fases i subfases del projecte. A continuació s'exposen les 8 fases en que s'ha separat el projecte, les fases compreses entre la 3 i la 7, ambdues incloses, no s'expliquen ja que venen imposades per la metodologia proposada.

Fases del projecte

- **FASE 1**. **Estudi i definició**. Es defineixen els objectius, es delimita l'abast del projecte i s'estudia l'estat de l'art del camp del coneixement en qüestió.
- **FASE 2**. **Desenvolupament de la metodologia**. S'adapta i desenvolupa una metodologia per tal d'acomplir amb els objectius i requeriments imposats.
- **FASE 3**. **Comprensió del negoci**.
- **FASE 4**. **Auditoria i comprensió de dades**.
- **FASE 5**. **Neteja, integració i preparació**.
- **FASE 6**. **Anàlisi preliminar**.
- **FASE 7**. **Avaluació, recomanacions i desplegament**.
- **FASE 8**. **Confecció de la memòria**. Es redacta i documenta tots els passos que s'han seguit per la realització del projecte, així com el resum i les conclusions.

	Setmanes																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
FASE 1 - Estudis i definició	1	2	3																
Definició del projecte i objectius	1																		
Estudi de l'estat de l'art	1	2	3																
FASE 2 - Desenvolupament metodologia			1	2	3														
FASE 3 - Comprensió del negoci					1	2	3												
Entrevistes						1	2												
Lectura documentació					1	2	3												
FASE 4 - Auditoria i comprensió de dades					1	2	3	4	5	6									
Obtenció de les dades					1	2	3	4											
Exploració de les dades						1	2	3	4										
Inventari de dades							1	2	3										
Resum de situació									1	2									
FASE 5 - Neteja, integració i preparació										1	2	3	4	5	6				
Neteja de les dades										1	2	3	4						
Enriquiment i millora												1	2						
Integració												1	2	3	3				
FASE 6 - Anàlisi preliminar																1	2	3	
Realització de diversos anàlisi																1	2	3	
Extracció de conclusions																1	2	3	
FASE 7 - Avaluació, recomanacions i desplegament																		1	2
Avaluació																		1	2
Recomanacions i desplegament																			1
FASE 8 - Confecció de la memòria					1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Taula 12. Diagrama de Grantt de la planificació del projecte

11.2. Costos associats

Els costos associats a la realització del projecte són molt baixos, ja que el projecte ha sigut realitzat per una sola persona i únicament s'ha fet ús d'un ordinador, software de tractament de dades gratuït i el paquet ofimàtic de Microsoft per a la realització dels informes i diagrames (tot i que es podria realitzar amb un software gratuït).

Cal tenir en compte que aquests costos són els associats a la realització del projecte, i no els costos associats al desplegament d'un entorn de processos analítics en tots els projectes de l'empresa. No s'ha tingut en compte el cost associat a les entrevistes realitzades, ja que quan una empresa es posa en mans d'una consultora, per rebre un servei d'aquestes característiques, ja accepta aquests costos indirectes.

SALARIAL	Cost horari	Personal	Temps	Cost
Estudiant	7,00 €	1	450 h	3.150,00 €
MATERIAL	**Inversió**	**Amortització**	**Ús**	**Cost**
Ordinador	1.200,00 €	1825 dies	133 dies	87,45 €
Paquet Ofimàtic	300,00 €	1825 dies	133 dies	21,86 €
			TOTAL	**3.259,32 €**

Taula 13. Costos associats a la realització del projecte

En la Taula 13 es poden veure els costos associats a la realització del projecte, que globalment sumen 3.260 €. Com a costos salarials, s'ha tingut en compte l'ajuda econòmica que rep un alumne en pràctiques de l'ETSEIB. Com a costos materials s'ha tingut en compte en, els dos casos, que s'amortitza la inversió en 5 anys, i que durant els dies que ha durat el projecte aquests béns no s'han usat per a cap altre fi.

Conclusions

En moltes organitzacions hi ha un volum de dades operatives molt elevat, en les quals és impossible realitzar-hi anàlisi perquè les dades estan optimitzades per la operativa del negoci i no per l'anàlisi, i per tant els manca una preparació prèvia que és totalment necessària. Anàlogament tampoc es poden realitzar molts anàlisi degut a la falta d'integració de les dades És a partir d'aquesta situació on els projectes de Data Mining i Data Warehousing, guanyen rellevància. El problema és que aquests tipus de projectes solen tenir uns costos associats elevats, impossibles d'afrontar en una empresa petita. Realitzant aquest projecte ha quedat demostrat que és possible aplicar una metodologia d'utilitat per a empreses que no poden fer front a una elevada inversió en consultoria, mitjans o personal per dur a terme un procés d'auditoria i preparació de dades.

La metodologia s'ha desenvolupat i adaptat a les necessitats establertes pels requisits previs i a la tipologia d'empresa, el fet de no establir a priori uns objectius molt específics d'anàlisi i dur a terme un procés més exploratori no ha suposat un problema i ha donat bons resultats. La metodologia proposada ha sigut capaç d'acomplir amb els objectius i requeriments establerts. Al fer-ne la prova de concepte s'ha obtingut uns convincents resultats que avalen la seva implementació en els altres projectes de l'empresa, a la vegada ha demostrat ser flexible amb l'abast de la seva execució, que dependrà dels recursos que es destinin al realitzar els anàlisis.

El procés d'auditoria i preparació de dades que s'ha dut a terme ha deixat patent la importància d'establir una cultura de gestió de dades per poder realitzar processos analítics. També ha permès extreure informació interessant simplement entenent i netejant les dades, sense un anàlisi pròpiament dit, demostrant la importància d'aquests processos a part de generar un taula d'anàlisi.

S' ha demostrat la viabilitat de crear i mantenir un entorn d'anàlisi utilitzant eines de *software lliure* i amb recursos limitats tant econòmics com de personal. El personal que gestionaria l'entorn seria de l'empresa, amb la possibilitat d'alguna contractació esporàdica en cas de voler aplicar tècniques de modelatge avançades. El *software free and open source* usat ha demostrat ser totalment vàlid i a la vegada ser flexible i modular. Aquestes característiques fan que les eines utilitzades siguin recomanables per a organitzacions interessades en l'anàlisi, però limitades pel preu d'entrada que suposa un software comercial.

Com a treball futur es podria millorar l'entorn implementant indicadors o estudis més automatitzats d'anàlisi que es considerin interessants per aplicar regularment. Tanmateix dur a terme modelatge fent ús de tècniques de Data Mining més avançades.

Agraïments

A Enertika i els seus treballadors per donar-me accés total a les dades del projecte, llibertat per fer-ne ús i paciència per resoldre els meus dubtes.

A Lluís Talavera, tutor del treball de fi de grau, per l'orientació i suport prestat durant la trajectòria del projecte.

A la meva família per ser una de les peces més importants per haver pogut arribar fins aquí sense defallir.

Bibliografia

[1] HAN, Jiawei. Data Mining. LIU, LING and ÖZSU, M. TAMER (eds.), *Encyclopedia of Database Systems*. Boston, MA : Springer US, 2009. ISBN 978-0-387-35544-3.

[2] KRIEGEL, Hans-Peter and SCHUBERT, Matthias. KDD Pipeline. LIU, LING and ÖZSU, M. TAMER (eds.), *Encyclopedia of Database Systems* [online]. Boston, MA : Springer US, 2009. [Accessed 19 December 2014]. ISBN 978-0-387-35544-3. Available from: http://www.springerlink.com/index/10.1007/978-0-387-39940-9

[3] FAYYAD, Usama, PIATETSKY-SHAPIRO, Gregory and SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. *AI Magazine* [online]. 15 March 1996. Vol. 17, no. 3, p. 37. [Accessed 28 April 2016]. DOI 10.1609/aimag.v17i3.1230. Available from: http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230

[4] CHAPMAN (NCR), Pete, CLINTON (SPSS), Julian, KERBER (NCR), Randy, KHABAZA (SPSS), Thomas, REINARTZ (DAIMLERCHRYSLER), Thomas, SHEARER (SPSS), Colin and WIRTH (DAIMLERCHRYSLER), Rüdiger. *IBM SPSS Modeler CRISP-DM Guide* [online]. 2000. [Accessed 10 March 2016]. Available from: https://the-modeling-agency.com/crisp-dm.pdf

[5] PIATETSKY, Gregory. CRISP-DM, still the top methodology for analytics, data mining, or data science projects. *KDnuggets* [online]. 2014. [Accessed 25 April 2016]. Available from: http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

[6] SONG, Il-Yeol. Data Warehouse. LIU, LING and ÖZSU, M. TAMER (eds.), *Encyclopedia of Database Systems*. Boston, MA : Springer US, 2009. ISBN 978-0-387-35544-3.

[7] WALLS, David and SCOTT, Mark D. 7 Steps to Data Warehousing. *SQL Server Pro* [online]. 1999. [Accessed 15 April 2016]. Available from: http://sqlmag.com/database-administration/7-steps-data-warehousing

[8] BOIRE, Richard. Data Mining: The Data Discovery. *Richard Boire Blogspot* [online]. 2012. [Accessed 10 March 2016]. Available from: http://richard-boire.blogspot.com/2012/11/the-data-discovery-investing-in.html

[9] SATTLER, Kai-Uweh. Data Quality Dimensions. LIU, LING and ÖZSU, M. TAMER (eds.), *Encyclopedia of Database Systems*. Boston, MA : Springer US, 2009. ISBN 978-0-387-35544-3.

[10] Data integration. *Wikipedia* [online]. [Accessed 20 April 2016]. Available from: https://en.wikipedia.org/wiki/Data_integration

[11] Semantic integration. *Wikipedia* [online]. [Accessed 20 April 2016]. Available from: https://en.wikipedia.org/wiki/Semantic_integration

[12] MIKE, Bergman. Sources and Classification of Semantic Heterogeneities. *AI3,*

Adaptative Information [online]. 2006. [Accessed 11 May 2016]. Available from: http://www.mkbergman.com/232/sources-and-classification-of-semantic-heterogeneities/

[13] PYLE, Dorian. *Data Preparation for Data Mining* [online]. Morgan Kaufmann Publishers Inc., 1999. [Accessed 27 February 2016]. ISBN 1558605290, 9781558605299. Available from: http://dl.acm.org/citation.cfm?id=299577

[14] What is an ESCO? *NAESCO (National Association of Energy Service Companies)* [online]. [Accessed 16 April 2016]. Available from: http://www.naesco.org/what-is-an-esco

[15] Energy Service Companies | Energy Efficiency. *European Commission, Institute for Energy and Tranport* [online]. 2016. [Accessed 16 April 2016]. Available from: http://iet.jrc.ec.europa.eu/energyefficiency/esco

[16] PUIG, Xavier, ALCALÁ, Victor and YAÑEZ, Josep. *Memoria Técnica Free-Cooling*. Barcelona, 2014.

[17] *International Performance Measurement and Verification Protocol: Concepts and Options for Determining Energy and Water Savings*. 2002.

[18] FREITAS, Luis and PUIG, Xavier. *Protocolo de Medida y Verificación Telefonica MX*. Barcelona, 2014.

[19] SCHNEIDER, Adam. GPS Visualizer: Assign elevation data to coordinates. [online]. [Accessed 17 May 2016]. Available from: http://www.gpsvisualizer.com/elevation

[20] PETRAGLIA, Antonio, SPAGNUOLO, Antonio, VETROMILE, Carmela, D'ONOFRIO, Antonio and LUBRITTO, Carmine. Heat flows and energetic behavior of a telecommunication radio base station. *Energy* [online]. September 2015. Vol. 89, p. 75–83. [Accessed 4 February 2016]. DOI 10.1016/j.energy.2015.07.044. Available from: http://www.sciencedirect.com/science/article/pii/S0360544215009391

[21] SERGIO SIBILIO, LUIGI MAFFEI, Raffaello Possidente. 566: Thermal and energetic analysis of a precast panel for industrial buildings. In : *25th Conference on Passive and Low Energy Energy Architecture (Dublin 2008)* [online]. 2008. [Accessed 17 May 2016]. Available from: http://citeseerx.ist.psu.edu/viewdoc/citations?doi=10.1.1.521.6369

[22] FORSYTHE, William C., RYKIEL, Edward J., STAHL, Randal S., WU, Hsin-i and SCHOOLFIELD, Robert M. A model comparison for daylength as a function of latitude and day of year. *Ecological Modelling* [online]. June 1995. Vol. 80, no. 1, p. 87–95. [Accessed 17 May 2016]. DOI 10.1016/0304-3800(94)00034-F. Available from: http://www.sciencedirect.com/science/article/pii/030438009400034F

[23] Weather Forecast & Reports, Weather Underground. [online]. [Accessed 18 May 2016]. Available from: https://www.wunderground.com/

[24] THE R FOUNDATION. R: What is R? *R Project* [online]. [Accessed 26 May 2016]. Available from: https://www.r-project.org/about.html

[25] Pandas: Python Data Analysis Library. *Pandas* [online]. [Accessed 26 May 2016]. Available from: http://pandas.pydata.org/index.html

[26] TALEND. Talend Open Studio: Open Source ETL & Data Integration. [online]. [Accessed 26 May 2016]. Available from: https://www.talend.com/products/talend-open-studio

[27] Anaconda. *Continuum Analytics* [online]. [Accessed 13 June 2016]. Available from: https://www.continuum.io/why-anaconda

[28] Matplotlib. *Matplotlib* [online]. [Accessed 2 June 2016]. Available from: http://matplotlib.org/